Abstracts

Business Applications
Advance your company's data mining power – make better decisions and predict new opportunities. Learn how innovative companies are using data mining to make strategic business decisions in the areas of business intelligence, database marketing, performance management, compliance and data analysis.

Perspectives from the Financial Services Industry
Data mining helps financial institutions make sense of the vast amounts of data they collect. Explore the different ways financial companies, insurance companies and banks use data mining to determine market and industry characteristics, predict individual company and stock performance, manage anti-money laundering strategies, reduce risk and increase customer retention.

Fraud Detection
Fraud detection is of great interest to many companies and organizations. The aggregate totals for fraud losses are enormous, but with the use of data mining, learn how companies can identify potentially fraudulent transactions and contain the damage.

Perspectives from the Healthcare Industry
Data mining can help predict the effectiveness of surgical procedures, diagnostic tests, medications, service management and process control. Discover how healthcare organizations use data mining to improve patient care while minimizing costs.

New and Emerging Technologies
Acquire valuable perspectives on the latest advancements in data mining technologies and techniques. Hear industry-renowned speakers discuss ground-breaking research and innovative new methods for using your data.

Perspectives from the Retail Industry
In today's market, generating retail sales requires selling the right product to the right customer at the right time ? every time. Learn how retail companies use data mining to remain profitable through effective customer segmentation, intelligent merchandising, targeted marketing campaigns, efficient supply chains and optimized working capital.

Data Mining for Marketing
A marketing organization's effectiveness is directly impacted by its ability to leverage customer/market insight through the usage and benefits of data mining. Learn how to use your data to predict customer behavior, achieve greater customer acquisition and retention, segment your customers for better understanding, determine the incremental profitability of new products, conduct targeted campaigns to qualified audiences, and grow your market share.

The following abstracts have been provided:

Linear models, colinearity and Variable Selection for giga bases
Leonardo Auslender, SAS
Variable selection is heavily used in the process of model building with large giga bases of observational data. While variable selection methods, typically of the stepwise family, are (or should be) fairly well understood, issues such as multicolinearity become the bete noir especially when the resulting model search it not very successful. We analyze the merits of these claims, and propose a user-controlled tool that remedies colinearity resulting of the variable selection search.

Data Mining - Interactive Marketing
Kim Bell, Data Analysis, Inc.
In the high-speed world of internet advertising, business owners are realizing that one of the most effective means of increasing online market share is through a combination of search engine optimization and pay-per-click advertising. Finding a way to help customers

find you is the key to successful interactive marketing campaigns. Through data mining techniques, analyses are conducted on such categories as: keyword searches, visitor loyalty, visitor recency, search engine sources, website visits, page views, geographical locations, cost per clicks, advertised words (adwords), adword positions, and conversions from website visits to revenue dollars. Information used in the analyses includes, but is not limited to,

- the total number of visits and page views your site received
- the average number of page views per visit (P/V)
- the number of visits and page views over time (Averages are calculated over the entire selected date range including dates not yet elapsed when applicable)
- the number of first-time visits and returning visits
- the cities from which the most visitors come to your site
- top referral sources
- impressions (the number of times an ad was shown)
- clicks (the number of times an ad was clicked)
- trans (the number of transactions that occurred after a referral from this Source[Medium] or keyword)
- cost for internet advertising.

This session will discuss the means of collaborating each of the above data points to increase website traffic, generate new business, lower customer acquisition costs, and enhance product awareness throughout the online marketplace.

Financial Data Mining: What to do with the money you bring home from Vegas

Gary D. Boetticher, University of Houston - Clear Lake

Financial data mining has been described as "the toughest way to make easy money." There are many traps and misconceptions about financial modeling. Commercial products exist which exaggerate their findings or make faulty assumptions regarding their modeling process. Academic research may forget to discount "buy and hold" strategies in their results.

This session examines how to apply genetic programs to the stock market and optimize financial model building.

- It describes various traps in financial data mining process and offers different techniques for building sound financial models.
- It examines how to extend genetic programs for financial data mining.
- It describes different aspects of financial domain knowledge and how to incorporate it into the financial data mining process.
- It offers demonstrations on building financial models using stock data and/or future contracts.

Predictive Modeling and Segmentation in Newspaper Industry

Goutam Chakraborty, Ph.D., Associate Professor of Marketing, Oklahoma State University

In this talk, I will discuss an application of predictive modeling and segmentation for a major newspaper in the southwestern United States. The data environment includes both address-specific (individual household level) data as well as grouped (census block group level) data. Modeling task includes predicting subscription (yes/no) as well as the type of subscription (Regular including Sunday, Regular but no Sunday, 3-day weekend and Sunday-only) for households.

Using administrative records to assess the performance of health care providers - Pitfalls and Challenges

HungChing Chan, Medica

Health insurance claims capture the majority of health services transactions, and provide a potential data source to improve quality and reduce costs of health care. Several case studies will be presented to demonstrate how insurance claims are used for these purposes. The challenges of mining health insurance transaction data will be discussed as well.

New Mathematical Optimization Functionality in SAS/OR

Manoj Chari, SAS Institute

Mathematical optimization is an analytical methodology that is closely associated with statistics and data mining - many predictive modeling and estimation procedures involve solving optimization problems, and many practical optimization based decision problems are based on input derived from statistical and data mining methods. In this presentation, we will discuss a completely new set of mathematical optimization procedures and solvers in SAS/OR whose initial release will be completed with SAS 9.2. These include

- PROC OPTLP for linear programming
- PROC OPTQP for quadratic programming with linear constraints
- PROC OPTMILP for mixed integer programming
- Two general nonlinear programming solvers based on sequential quadratic programming and interior point algorithms
- PROC OPTMODEL, a powerful algebraic modeling language with internal access to all of the above solvers.

Collectively these new procedures represent a quantum leap in functionality, performance,

flexibility, and ease of use for SAS/OR optimization.

Using business and statistically oriented examples, we will present a unified view of this new optimization functionality through PROC OPTMODEL. This procedure combines a high level, flexible programming framework with algebraic modeling capabilities and an optimization oriented syntax. We will show how the mathematical formulation of an optimization problem that a modeler writes down on paper can be passed virtually unchanged into OPTMODEL code. The procedure has an executable "solve" statement which can be used to invoke any of the above solvers. It is also possible, using the programming statements in PROC OPTMODEL, to implement customized optimization algorithms and heuristics that involve the iterative use of one or more types of optimization solvers.

Bayesian Methods in Asset Management and Risk Management

Joe Chen, Ph.D., Vice President Risk Modeling, Origen Financial

This presentation will include two parts:

1. Black-Litterman model is widely applied by practitioners in sophisticated asset allocations. The model uses a Bayesian approach to combine investor's subjective beliefs on the expected returns of assets with the market equilibrium returns (the prior). Bayesian analysis provides a mechanism to synthesize subjective views with empirical realities. The Black-Litterman model resolves major issues of Markowitz's classical mean-variance paradigm. The latter is very sensitive to the covariance matrix and error-maximizing, and often ends up with unintuitive, highly-concentrated portfolios. Based on the posterior distribution of the new returns, the Black-Litterman model leads to intuitive portfolios with reasonable portfolio weights. Extension can be made if the normality assumptions are relaxed in the model.
2. Under Basel II, banks can use their default probability estimates for calculating regulatory capital. Financial institutions use credit scoring models to underwrite and price loans. In either case, Bayesian inference can improve the accuracy of the default probability estimates. Bayesian methods also provide a natural way to handle structural differences between a bank's internal data and external data.

Bringing Data Mining Down to Business

Randy Collica, HP

Anyone who has performed or been involved with data mining and business intelligence will probably recognize that one of the first questions to be answered is: "What is the main business problem or issue that needs to be solved?" Too often management's expectation of data mining is "help me find something interesting" or "what can you do with my data." In all of these situations, translating the business problem or issue into a form that can be addressed by data mining or modeling is the first order of business. Then, setting expectations for all aspects of the project will go a long way at ensuring that there are no real "surprises" in the results. Understanding the main business problem should precede any data mining or modeling and having this understanding at the outset will increase the likelihood of success in the overall project.

This presentation will look at several case studies in data mining and business intelligence projects and their true business issue or pain. Key findings and experiences will be shared with examples in survival data mining, forecasting, structured data mining, and text mining. Another key learning that will be addressed is the how to explain results from complex models and analyses to an audience that has little experience with data, statistics, or modeling.

Using Data Mining to Build an Effective Recruitment and Retention Strategy

Cali M. Davis, J. Michael Hardin, The University of Alabama, Tom Bohannon, Baylor University

Freshmen recruitment, enrollment, and retention are important issues in higher education. Researchers from across the country have used data mining techniques to develop models to predict who will enroll and why students fail to enroll for their second year. Over the past several years, the University of Alabama, a public university, and Baylor University, a private university, have used SAS Enterprise Miner and data mining to develop predictive models from admissions and records data.

In this presentation, the data mining experiences of both of these universities will be discussed, as well as a comparison and contrast of the private and public university perspective. In addition, other applications of data mining in higher education such as donor modeling and student recruitment strategies will be discussed.

Building a Data Mining Community

David Duling, SAS

Data miners typically work in a very competitive environment. Statistical models that impact business success are treated as corporate assets that are subject to tight security. This includes the business purpose, data selection, target definition, feature aggregation, and model technique. Data miners often acquire techniques from vendors and ideas from conferences, journals, and informal discussions. Enterprise Miner extension nodes provide a mechanism for sharing basic functions between users in an organization, between consultants and customers, and ultimately between users across organizations. This

presentation discusses the extension node mechanism, presents examples of nodes that have proven to be useful, and proposes a program for exchange between users.

Using SAS Data Mining Procedures to estimate the 'true' value of new (and existing) products and features

Abdu Elnagheeb, Bank of America

One of the more difficult analytic challenges faced by financial and other industries is to determine the incremental profitability of products that are rushed to market without the benefit of a well thought out experimental design. It is not uncommon for a product or feature to be rolled out across the franchise, and then have senior management ask: "How much incremental profit is being generated?"

Too often product managers will report that customers buying the new product are worth 'X', customers buying the old product are worth 'Y', so the incremental value of the product is X - Y. But common sense and experience tell us that the customers selecting X over Y maybe be fundamentally different customers, and that self-selection, differential opportunity, and other inherent biases invalidate such analyses.

To overcome the influence of these biases, we have developed a methodology that uses CHAID, Linear, and other modeling techniques to back out customer and market influences, and more precisely zero in on the true value of the product.

This presentation follows the development of the methodology using artificial as well real data and actual products to demonstrate how the true value of a product can be estimated, even when there are strong biases inherent in the purchase and use of the product. Variable selection, model optimization, actual code, and the shortcomings of the mythology are presented.

From Data Mining Grand Challenges to the New Sciences Underlying the Internet

Usama Fayyad, Yahoo! Inc.

As the Internet continues to change the way we live, find information, communicate, and do business, it has also been taking on a dramatically increasing role in marketing and advertising. Unlike any prior mass medium, the Internet is a unique medium when it comes to interactivity and offers ability to target and program messaging at the individual level. Coupled with its uniqueness in the richness of the data that is available, in the variety of ways to utilize the data, and in the great dependence of effective marketing on applications that are heavily data-driven, makes data mining and statistical data analysis, modeling, and reporting an essential mission-critical part of running the business. However, because of its novelty and the scale of data sets involved, few companies have figured out how to properly make use of this data.

In this talk, I will review some of the challenges and opportunities in the utilization of data to drive this new generation of marketing systems. I will provide several examples of how data is utilized in critical ways to drive some of these capabilities. The discussion will be framed with the more general framework of Grand Challenges for data mining: pragmatic and technical.

I will conclude this presentation with a consideration of the larger issues surrounding the Internet as a technology that is ubiquitous in our lives, yet one where very little is understood, at the scientific level, in defining and understanding many of the basics the Internet enables: Community, Personalization, and the new Microeconomics of the web. This leads to an overview of the new Yahoo! Research and its aims: inventing the new sciences underlying what we do on the Internet, focusing on areas that have received little attention in the traditional academic circles.

Visualizing Multiple and Logistic Regression Models

George Fernandez, University of Nevada, Reno

Data mining techniques stress visualization to study thoroughly the structure of data and to check the validity of statistical model fit to the data. Multiple and binary logisitic regression are widely used supervised learning methods in data mining. In multiple linear regression models, problems arise when serious multicollinearity or influential outliers are present in the data. Failure to include significant quadratic or cross-product terms results in model specification error. Simple scatter plots and logit plots are not effective in revealing the complex relationships of predictor variables or data problems in multiple linear and in logistic regression. In multiple linear regression applications, partial regression plots are considered useful in detecting influential observations and multiple outliers; partial residual plots or the added-variable or component-plus-residual plots are useful in detecting non-linearity and model specification errors. The leverage plots are considered effective in detecting multicollinearity and outliers. The VIF-plot, which is very effective in detecting multicollinearity, can be obtained by overlaying both partial regression and partial residual plots with a common centered X-axis. In case of binary logistic regression (BLR), the partial delta logit plots are useful in detecting, significant predictors, non-linearity, and multicollinearity. The partial delta logit plot illustrate the effects of a given continuous predictor variable after adjusting for all other predictor variables on the change in the logit estimate when the variable in question is dropped from the BLR. By overlaying the simple logit and partial delta logit plots, many features of the BLR could be revealed. User-friendly SAS macro applications for generating these exploratory plots will be presented.

Operationally Significant Patterns of Association

William C. Hardy and Richard W. La Valley, SAIC, Inc. Advanced Systems Concepts

An association between two persons is an observation or reported fact that can reasonably be assumed to indicate they know of, and probably communicate with, each other. Much of the information on known/suspected terrorists comprises collections of such associations absent information that would support determination of whether any particular set of associations evinces cooperative involvement in pursuit of a common goal or objective. A major challenge for information technology is to develop tools for discovering such stronger relationships in large collections of associations gathered from disparate sources. To this end, this paper describes patterns of associations that are indicators of possible organizational activity, regardless of the means of their discovery. It describes for each:

- techniques for visualizing the pattern in the data;
- efficient means of discovery of instances of the pattern in very large link data bases; and
- procedures for validating instances of the pattern as being attributable to organizational behavior rather than a chance encounters.

It further describes results of analyses of real-world data that demonstrate the operational utility of inferences derived from the patterns.

Analytical Roadmap – the Marketer's Map to Improved ROI

Maria Marsala Herlihy, KnowledgeBase Marketing, Inc.

If you're ready to take your marketing campaigns to the next level, this session is for you! Many marketers want to employ analytics but don't know where to start. This session introduces you to the Analytical Roadmap AND you won't need a statistics degree to comprehend. All the concepts are presented in terms that you will understand illustrated with real-life case studies from companies that benefited from using analytical tools. This is a must-attend session for marketing professionals looking to improve their ROI!

You'll learn

- complex analytical concepts in familiar terms that you can understand
- when to use analytics, how to get started and which tools are appropriate for your marketing challenges.
- how to bridge the gap between traditional direct marketing and database marketing
- practical examples of how to apply analytics for increased ROI.

Training Health Care Professionals in Data Mining

Mark E. Johnson & Morgan Wang, University of Central Florida

Since M2005, we completed the delivery of a 5 session (3 days per session) data mining short course for a major health care insurance company. This course was primarily geared for marketing and information technology professionals who had considerable experience with SAS but limited exposure to SAS Enterprise Miner. Highlights of the training will be described including the incorporation of company specific projects and carefully selected pedagogical data sets. Guidelines on the balance among theory, application and output interpretation will be given. Finally, comparisons and contrasts of this data mining training experience to other training in which the presenters have been involved (data mining, black belt six sigma training and design for six sigma) will be provided.

"Supervised" Customer Segmentation Analysis using SAS Enterprise Miner 5.2

Paul B. (Brad) Jordan, Leader, Marketing Analytics, Blue Cross Blue Shield of Florida

Customer segmentation is one important way to increase marketing efficiency. Traditionally, clustering algorithms are used this goal. However, clustering techniques are unsupervised and there is not any easy way to assess the clustering performance. Using a "supervised" data mining approach has helped us achieve the desired result of producing segments that are linked to a target variable.

What's Leaving Your Wallet? Estimating Industrial Share of Spending in a Competitive Environment with Incomplete Information

Talbot Michael Katz, Analytic Data Information Technology Consultant

A holistic view of a customer base is a powerful weapon in the hunt for consumer dollars, but we typically confront several holes in our efforts to construct such a portrait. We are almost always missing a great deal of information about the full spending profile of each customer, and the customers for which we have the best data may not be representative of the entire base. And although spending in separate industries may be influenced by unrelated factors and have very different characteristics, because all dollars are in competition, it may be desirable to treat the entire spectrum of industries in concert, especially to be able to consider the key element of elasticity. We will examine how to piece together a picture of wallet share using classic methodologies such as segmentation, discrete choice modeling, and seemingly unrelated regressions, keeping in mind such issues as handling of missing data and variable selection. And we will look at promising new approaches to explicitly exploit the gaps in information, such as Chen and Steckel's hierarchical Bayesian analysis of interpurchase times.

Detecting Fraud by Building Strong Control Environments using Continuous Monitoring

William (Bill) J.Kelley, COL FA, Program Director, Data Mining Division, Department of Defense, Office of the Inspector General
Some topics to cover include

- establish effective review and approval processes
- utilize human capital management to foster accountability
- implement innovative monitoring procedures to eliminate fraud and abuse by applying business rules that identify transactions with high risk

Justification:
With the continued downsizing of oversight resources, governmental agencies and the private sector need to take advantage of continuous monitoring techniques to identify transactions that have high risk. Although the DoD has taken aggressive action to improve its programs, most programs still need better implementation and oversight of management controls at the corporate and activity level.

Needs based segmentation and its application in SKTelcom

Jongdo Kim, SKTelecom, Korea
As services and products grow various and complicated, it is harder for customers to find right service for themselves and for company to find right target. In this session, how SKTelecom is segmenting customers and using segmentation to offer services will be discussed. We segmented customers based on their needs, that we assume to be revealed by their behaviors. We are using this segmentation to increase sales and to enhance our customer satisfaction.

Word of Mouth and Opinion Spreading: New Sociophysics Approaches

Dmitri Kuznetsov, Ph.D., Senior Brand Analyst, Media Planning Group
Understanding of leading mechanisms of opinion spreading in social systems is a key point for powerful marketing and political campaigns. Opinion exchange and convincing (so-called, "word of mouth") is a crucial part of the opinion spreading along with advertising, experience, and economic and political factors. During last decade opinion dynamics became one of the central parts of the scientific branch named Sociophysics that applies modern methods of statistical physics to social phenomena. Its success originates from the deep similarities in rising (1) from single-particle properties to complex-object features in physics and (2) from single person characteristics to social mass behavior. In 2005 we introduced Mediaphysics as a part of Sociophysics, studying processes of mass communications in social systems and demonstrated its potential for applications in different areas. In this presentation, different sociophysics approaches to the "word of mouth" and opinion spreading are demonstrated and discussed.

Practical and Flexible Modeling with the GNBC: A Case Study

Kim Larsen, Director, Database & Relationship Marketing - DMSA, Charles Schwab & Co., Inc.
Predicting the probability of a binary event given a set of independent variables is one of the most commonly occurring modeling problems, whether it's risk management, marketing, or a medical study. Last year at M2005, I presented the Generalized Naïve Bayes Classifier (GNBC) and argued that it's one of the most powerful tools for this task. In short, the GNBC is a generalization of the well-known Naïve Bayes Classifier that

- treats input variables in a flexible nonparametric fashion
- uncovers hidden patterns in the data
- can handle problems of large dimensions.

This year I'll re-visit the GNBC, but spend less time on theoretical background and justification, and instead concentrate the talk around a case study using real-life data. The case study will demonstrate

- how to use, interpret, and implement the GNBC
- the power of the GNBC and how it compares to other well-known techniques.

Customer Acquisition and Retention using Data Mining Techniques

T. Lynn Locke, Director, Database Marketing, Blue Cross Blue Shield of Florida
A complete case study that illustrates how to utilize data mining techniques to enhance the customer acquisition process will be presented. The resulting model from this study has been tested in January and April 2006. Utilizing this model, the marketing department can afford to mail four times more volume than in prior years, while reducing the per customer acquisition cost by 67%.

Why Data Mining Is an Encapsulating Solution to Querying, BI and BA in Research and Planning in Higher Education and Beyond

Jing Luan, Educational Services and Planning, San Mateo, CCCD
The session covers the general use of data mining applications for higher education research and decision making with a focus on understanding data mining being the ultimate system approach to informatics. Specifically, since data mining resides on the top of the 3 tiered Knowledge Management model developed by Dr. Jing Luan, he will describe the use data mining applications as an end game for any of the three main purposes of data mining: data queries, report generation (BI), and predictive modeling (BA).

Preparing the Data Mining data: Techniques for harmonizing and integrating information from disparate datasets

James (Jim) W. Mentele, Central Michigan University - Research Center
This talk will focus on the activities of data mining projects typically requiring 80-90% of the project effort. Harmonizing and integrating data from multiple functions and departments, customers & vendors, governments and 3rd party data sources, over various time-periods and levels of granularity presents challenges. The author will discuss an overview of techniques with examples from 40 years of related experience.

Business Intelligence Success Factors - Essential Skills for Success in a High-tech, Data-driven Corporate World

C. Olivia Parr-Rud, OLIVIAGroup
Data mining and business intelligence solutions offer incredible opportunities for companies to improve their bottom line. However, increased complexity and the constant introduction of new technologies as well as shifting markets and consumer demands require us to master a broader set of skills. Expert leadership has never been more important. Studies show that roughly 70% of business intelligence projects fail due to poor leadership. This presentation will discuss concepts and provide insights into some basic skills that are essential for success in a high-tech, data-driven industry that is quickly going digital, virtual and global.

Communication is an essential skill in the complex world of business intelligence. As we turn to high tech solutions for our database marketing challenges, our employees require highly specialized and varied skill sets. In many organizations, each group speaks a "different language." Yet their ability to effectively communicate is critical to our business success.

Effective Collaboration is critical to leveraging highly specialized skill sets among various groups or entities. At every level of an organization, it is no longer possible to do it all. Mergers and acquisitions are creating additional challenges. For collaboration to succeed, leaders must have the ability to build an environment of trust while encouraging competition.

Creativity is quickly becoming the new competitive advantage. Business intelligence solutions are linear processes that can be automated or outsourced. Increasingly, these processes are becoming available to our competitors. We must become proficient at 'whole brain' thinking to creatively utilize these tools and insure continued innovation.

Managing Change requires vision and leadership. To leverage opportunities that are on the horizon, we must thrive on uncertainty, inspire our employees, trust our instincts and become intelligent risk takers. The rewards await those leaders who set clear goals, move forward with courage and trust the process.

Determining the Best Balance of In-House and Outsource for your Segmentation and Building Internal Analytic Competency

Will Neafsey, Ford Motor Company
As market research and brand organizations are constantly streamlined and rationalize with today's business environment, marketers are continually faced with the decision of outsourcing their analysis or doing it in-house. This can be a complex decision, particularly with segmentation. The application of analytic and segmentation techniques can be very industry specific. Valuable time and money can be wasted educating a consultant on your industry, your company and your consumers. On the flip side, hiring, training, staffing and protecting an analytic/segmentation team can be difficult. This session will discuss the advantages and pitfalls of bringing segmentation and analytic competency in-house.

An Epidemiological Framework for a Hypothesis Generating Investigation of Multiple Near?Concurrent Vaccinations and Potential Health Endpoints

Daniel C. Payne, PhD, MSPH[1], Charles E. Rose, PhD[1], Aaron Aranas, MPH, MBA[1], Michael M. McNeil, MD, MPH[1], Jodi Blomberg, MS[2]
Affiliations: [1] U.S. Centers for Disease Control and Prevention, National Immunization Program, Epidemiology and Surveillance Branch, Bacterial Vaccine-Preventable Diseases Branch, [2] SAS, Inc.

Limited scientific knowledge exists regarding whether health effects are associated with receiving multiple near-concurrent (MNC) vaccinations among adults. The medical literature available on this topic addresses primarily isolated observations following specific childhood vaccine combinations. To our knowledge, no defined framework for evaluating this topic in a comprehensive and standardized fashion using a large health outcomes dataset has been reported. A prototype framework using SAS Enterprise Miner applications to study this topic will be presented.

Staff from the Centers for Disease Control and Prevention (CDC) and SAS, Inc. applied SAS Enterprise Miner software to data collected by the Defense Medical Surveillance System (DMSS), a longitudinal surveillance database administered by the Army Medical Surveillance Activity. We used SAS Enterprise Miner to conduct a range of analyses, including decision tree modeling and sequence analysis, to define and describe this hypothesis generating framework using the full DMSS dataset of medical, vaccination, administrative, demographic, and deployment information for over 7 million individuals. Our approach enabled analysis of the topic of MNC vaccinations and health endpoints to be

viewed from 4 relevant perspectives of exposure: differing combinations of nominal vaccine types, live versus non-live vaccines, vaccines containing aluminum adjuvant versus those without this adjuvant, and lastly, by a continuous variable representing cumulative antigen counts.

Data Mining to Determine Characteristics of High-Cost Diabetics in a Medicaid Population

S. Greg Potts, MBA, Data Mining Team Leader, Arkansas Foundation for Medical Care, Office of Projects and Analysis

Diabetes Mellitus is a chronic condition affecting thousands of Arkansans. Arkansas Medicaid recipients with this condition often incur significant medical costs related to treatment of the condition.

Our objectives were twofold: 1) to use the Decision Tree algorithm in SAS Enterprise Miner 4.3 to determine high cost-drivers among Arkansas Medicaid recipients with Diabetes, and, 2) to determine whether the study cohort could be segmented into small, manageable groups for potential disease-management intervention in an effort to manage costs and provide quality care. Ours was a retrospective cohort study using all paid claims for recipients enrolled 12 continuous months during State Fiscal Year 2004 with a diagnosis of Diabetes Mellitus from Arkansas Medicaid's Decision Support System (DSS) Data Warehouse.

Elliptical Predictors in Logistic Regression

Will Potts, Capital One

Many customer attributes have nonnormal features such as skewness and heavy tails. Logistic regression does not involve distributional assumptions on the predictors. Nevertheless, normalizing transformations can make it easier to discover models that are both simple and powerful. An elliptically contoured predictor space sharpens the relationship between the response and explanatory variables by reducing nonlinear confounding and the influence of sparse regions. The benefits of predictor transformations are demonstrated using credit bureau data.

Exploring Open Source Software Development and Maintenance Using Data Mining and Text Mining

Uzma Raja, Department of Information Systems, Statistics, and Management Science, Culverhouse College of Commerce, University of Alabama

Experiences of an exploratory study of the Open Source Software development and maintenance will be shared in this talk. The corporate use of OSS is increasing and there is a need for evaluation models of these projects. In this research data mining and text mining techniques were used to discover knowledge from transactional datasets maintained on OSS projects. Models were formulated, tested and validated using these very large and comprehensive datasets. SAS Enterprise Miner and SAS Text Miner were used to develop the performance evaluation models.

In this presentation, the data mining and text mining results of the OSS projects will be discussed. The various factors that affect the outcomes of OSS projects will be highlighted. In addition some experiences using text mining to improve model performance will also be discussed.

Business Price Optimization using Data Mining and Modeling

Timothy D. Rey, The Dow Chemical Company

Businesses at Dow are always trying to set their prices correctly to obtain a reasonable margin for operating the business. This is a complex problem in that many issues can come into play: Raw Material Costs, Plant Upsets, Accounting Practices, Supply/Demand in the market, Customer buying patterns, to name a few. This project combined many data sets to study internal and external trends. SAS Enterprise Miner was initially used to the do the exploratory Data Mining. Eventually an iThink nonlinear simulation model was build for the business to use. The Dow Data Mining and Modeling process followed to conduct the project will be reviewed along with the interim results and final model.

Customer Profit Value in Insurance Business

Günter Schmölz, Uniqa Insurance Austria, Customer Intelligence

UNIQA is Austria's biggest insurance company, is active in more than 14 European countries and has more than 6 million customers. Uniqa has built up an international, comprehensive and standardized SAS-database architecture which enables an integrated data management process. This database is not only the basis for the data mining process with SAS Enterprise Miner but also the central basis for the whole campaign management, for the Client Reporting system, for all actuarial analyses and an important source for the SAS-Balanced-Score-Card.

This presentation shows how data mining and customer-data-management triggers Uniqa's client orientated strategy. The main results (individual client ratios) of the data mining and scoring process are implemented in the UNIQA client information system (which all employees have access to). This presentation will concentrate on the most important ratio, the Customer-Profit-Value. This Value is a prognosis of the expected client contribution margin for the next 12 months. The presentation will give an insight on how more than 50 customer-criteria are influencing the individual score. The Customer-Profit-Value enables

UNIQA to differentiate between good and bad customers and allows a very precise value prognosis not only on an individual basis but also on a segment level. This presentation will also show how UNIQA uses these customer orientated scores to steer the sales force.

Leveraging advanced statistical analysis to turn online marking data into actionable marketing insights

Jack Schwartz, CTO, [x+1]

Today, the online customer touch point is the life's blood for the bulk of consumer-focused companies. In order to maximize marketing performance it is imperative that the most robust, consumer centric analytics be applied to campaigns to ensure the optimal experience for existing customers, as well as prospective ones.

When users move through websites they leave imprints which tells you a little about themselves, their preferences on a particular product or service and willingness to buy. Tracking their movement allows a marketer to better tailor their campaign to get this user to the ultimate goal ? on-line purchase or repeat purchase.

But while the seasoned on-line marketer knows that this data will give them a competitive advantage, the primary challenge with web-analytics is to get them to act on the data. Now, as a growing number of companies return to data-driven marketing, some key techniques and principles need to be incorporated at the start of the online campaign to ensure solid results and sustainable growth.

A View of the Data Mining Revolution and Evolution

William B. Smith, Executive Director, American Statistical Association and Professor Emeritus of Statistics, Texas A&M University

With the advent of modern and more powerful computers with inexpensive memory and automatic measuring and collecting devices and techniques, the volume (n) of multidimensional (p) data available for 'analysis' has increased beyond belief. As a statistician, the speaker will trace the revolution/evolution by illustrating procedures appropriate for various combinations of n and p. Traditional statistical methods focused on problems assuming small n and small p. Small n and large p problems (commonly called 'the curse of dimensionality') have been attacked, also. However, given the massive information now available, the data mining community addresses 'large n, any p' problems. Examples of analyses will be given to illustrate the methodologies of the n and p combinations.

Mining Without a Hardhat: A New Approach to Safety at a National Laboratory

Judy Spomer, Sandia National Laboratories, Knowledge Discovery and Extraction

Sandia is a national security laboratory involved in a variety of research and development programs to help secure a peaceful and free world through technology. We develop technologies to sustain, modernize, and protect our nuclear arsenal, prevent the spread of weapons of mass destruction, defend against terrorism, protect our national infrastructures, ensure stable energy and water supplies, and provide new capabilities to our armed forces.

The widespread nature of activities conducted at Sandia presents a unique challenge: How to keep approximately 8,500 staff members safe and injury-free when the work involves radiological materials, explosives, chemicals, lasers, and so on?

In early 2004, the Environment, Safety, and Health organization at Sandia National Laboratories initiated an effort to develop a statistical model to identify characteristics that could forecast the likelihood of work-related injuries or illnesses by organization. The resulting Injury and Illness Predictive Model (IIPM) has been in production since December 2004. On a quarterly basis, current data is run through the model, generating a forecasted 6-month safety incident rate for organizations at the laboratories. Resources can then be focused on interventions for areas at risk.

This session will cover the approach, stumbling blocks, and methods used to develop this model in SAS as well as its deployment at Sandia.

Data Preparation for Analytics

Gerhard Svolba, SAS Austria

Data preparation for analytics is an essential task for successful analysis, data mining and predictions. The quality of the data mart and its preparation is one of the most important influence factors for model quality and analysis result.

Data preparation itself must not only be considered as a technical or coding task. Analytic data preparation includes also the business point of view of the underlying business question as well as analytical and data modeling considerations. The business question for example may impose restrictions on which periods may be used as input data. Different analytical approaches require different analytic data structures.

Nonetheless coding of a data mart is a very important task in data preparation including data extraction from source systems, table joins, aggregations and the creation of derived variables. Time saving and clever approaches for data preparation facilitate the data preparation phase and provide derived variables with high explanatory and predictive power.

The presentation will deal with a specific case study and show how data from various data sources can be made up to build a clever one-row-per-subject data mart.

This presentation shows mainly SAS Code from SAS Base (Datastep, Procedures, SAS Macro Language) further a few procedures from SAS/STAT and SAS/ETS are used.

Mining Medical Claims Data from Exploratory to Confirmatory Statistical Methods

Thomas T.H. Wan, Ph.D., M.H.S., Professor and Associate Dean for Research, College of Health and Public Affairs, University of Central Florida

Health services researchers have relatively concentrated in applying exploratory statistical methods in identifying the patterns of care and analyzing the variation in health services use. Despite the critical role medical claims data play in detecting billing errors or fraudulent behaviors of billing practice, little is known about the extent to which variations in frequency, types, and seriousness of deficiencies reflect actual differences in quality of care versus broader systemic differences in the utilization and provision of health services. In response, we propose to formulate an evidence-based approach to mining claims data for not only identifying patterns of billing errors, but also examining the relative influences of facility/organizational and contextual factors that may influence a specific pattern of Medicare/Medicaid fraud and abuse. Examples of applying exploratory and confirmatory statistical methods, using claims data, are presented.

Data Integration – Tools vs. Code. Tips and techniques for converting from user-written code to process workflows within SAS Data Integration

Cary White, Data Warehouse Team Leader, UNC-Chapel Hill

Many shops have a large investment in developer-written ETL code. In many cases, this code was written by a consultant or developer who is no longer employed by the company. When a decision is made to purchase an ETL tool, the company must also decide how much of this code can be reused within the new tool.

The SAS language is a especially powerful 4 GL language that is well-suited for data extraction, transformation and loading (ETL) In its V9 release SAS is now offering an industry-standard ETL tool now called SAS Data Integration.

This talk will present a case study of a conversion project in which ETL processes written in Base SAS were converted to process workflows using SAS Data Integration Studio.

Some of the topics covered will include

- best practices to apply in the conversion process
- what to keep vs. what to convert
- relationship of data quality to the data integration process
- creating reusable processes
- maintenance and Metadata – two of the most important reasons to use a tool
- good features/shortcomings of the tool.

Data Mining at the Texas State Auditor's Office

Tom Winn, Texas State Auditor's Office

In general, data mining is the iterative and interactive process of finding new and potentially useful knowledge from large quantities of data. The author's use of that term encompasses sophisticated statistical techniques, as well as many simpler data analysis methods. The mission of the Texas State Auditor's Office is to actively provide government leaders with useful information that improves accountability. This expository presentation will describe some data mining work at the Texas State Auditor's office in support of auditing and investigations. Particular attention will be given to fraud detection and risk assessment in the public sector.

Data Preparation for Insurance Fraud Detection

Terry Woodfield, SAS

Detecting fraudulent claims in property and casualty insurance requires not only good data, but domain expertise in deriving inputs for fraud modeling. Many of the variables available from an insurance company's database are categorical, such as gender, ICD-9 codes, employment classifications, accident codes, and body part codes. Diagnostic codes like ICD-9 codes have very high cardinality and require either automatic or manual grouping. Strategies for converting categorical variables into predictive modeling inputs will be described. Converting transactional data, such as medical payments, into predictive modeling inputs will also be discussed.

They asked for a Segmentation scheme, not clusters

Jeff Zeanah, Z Solutions, Inc.

A common business request is to develop a segmentation scheme to assist in understanding customers and thus approach the market. The qualitative analyst generally converts these requests to some form of clustering using available data, because this is what the analyst knows how to do. The restrictions of clustering approaches may have little to do with the actual business request leading to frustration for all parties. Using Genetic Algorithm Approaches in PROC GA in SAS/OR (experimental in 9.1.3, with expected release in 9.2) the analyst may optimize a segmentation scheme on any need

that can be quantified. This presentation presents a simple application to introduce the concepts of Genetic Algorithms and initial approaches to accomplish an implementation in PROC GA.