ABSTRACTS CORRESPONDING TO THE PRESENTATIONS INCLUDED ON THE CD

CONTENTS:

Dear Sir or Madam,

We encourage you to speak at our future conferences. We are especially interested in presentations with an industry focus and a practical emphasis. For more information related to presenting at our conferences, please contact Lisa Solomon (Email: lisas@salford-systems.com or phone 619.543.8880 x14).

If you wish to be added to our mailing list for regular conference updates or if you have questions related to the conference proceedings, please contact Stefanie Frederick (Email stef@salford-systems.com or phone 619.543.8880 x28).

Thank you for your interest in our conference proceedings. We look forward to seeing you at our future conferences.

Sincerely,

Lisa Solomon and Stefanie Frederick
http://www.salford-systems.com
http://www.cartdatamining.com
http://www.salforddatamining.com

GENERAL INTEREST PRESENTATIONS

John Elder, Elder Research, Inc.

"Avoiding the Top 10 Mistakes in Data Mining"

Data Miners discover key patterns in historical data to make effective decisions today. Their interdisciplinary techniques constitute a "crystal ball" -- which has enhanced performance in noisy, data-rich fields ranging from the stock market to credit risk assessment, and marketing to fraud detection. But, Data Mining is still as much an art as a science, providing many convenient ways to do wrong things with one's data. Case studies of (often personal) errors -- both simple and complex -- will be drawn from real-world consulting engagements. Best Practices for Data Mining will be (accidentally) illuminated by their (rarely described) opposites. These common errors range from allowing anachronistic variables into the pool of candidate inputs, to subtly inflating results through early up-sampling. You'll hear cautionary tales of endangered projects and embarrassed teams-- and, thereby perhaps, avoid such a fate yourself.

BIOMEDICAL (INCLUDING DRUG DISCOVERY) PRESENTATIONS

Joseph Cappelleri, Pfizer

"Using CART to Develop a Diagnostic Tool for Erectile Dysfunction"

Objectives: To create an abridged 5-item version of the 15-item International Index of Erectile Function (IIEF-5) as a diagnostic tool to discriminate between men with and without erectile dysfunction (ED), and to develop a clinically meaningful gradient of severity for ED.

Methods: 1152 men (1036 with ED, 116 without ED) who reported attempting sexual activity were evaluated using baseline data from four clinical trials of VIAGRA™ (sildenafil citrate) and two control samples. The statistical program Classification and Regression Trees (CART) was applied to determine optimal cutoff scores on the IIEF-5 (range, 5-25 if patient engaged in sexual activity; 1-25 if patient had an opportunity but no desire for sexual activity) to distinguish between men with and without ED, and to determine levels of ED severity on the IIEF-5 using the IIEF item on penetration frequency.

Results: The optimal cutoff score was 21 with men scoring less than or equal to 21 classified as having ED and those scoring above 21 as not having ED (sensitivity=0.98, specificity=0.88). The severity of ED was classified into five categories: no ED (IIEF-5 score, 22-25), mild (17-21), mild to moderate (12-16), moderate (8-11), and severe (5-7 if patient attempted sexual activity; 1-7 if patient had an opportunity but no desire for sexual activity). Results were validated in a separate, independent study.

Conclusions: The use of CART was instrumental in the application of creating a diagnostic tool that not only distinguishes between men with and without ED but also classifies levels of ED severity.

Marsha Wilcox, Boston University School of Medicine

"Using CART to Discern Models in Genetics: Alcoholism, Alzheimer Disease and Aging"

Recursive Partitioning (RP) as implemented in CART has been useful in discerning genetic models in Alzheimer Disease (AD), alcoholism and aging.

In AD, the discovery of genes associated with the disease brings the hope of understanding the disease process and, ultimately, new therapies for prevention and cure. It is important to understand how these genes function – either alone or in concert with other genes. We found an association with a novel gene with known function and AD. The question arose about an interaction of the new gene with a well-established causal gene. Was the new gene simply present with the established gene, and thereby a false positive with no new contribution to the disease process; was it interacting with the established gene, and not contributing to the etiology of the disease on its own; or does it have an independent influence on the occurrence of disease? RP helped show that the new gene is responsible for some cases of AD.

Alcoholism is a complex disorder with complex etiologies. It is clear that, for some people, the liability toward alcohol dependence is inherited. We examined a genome screen (~400 genetic markers) for linkage and association with the disorder. We found several interesting loci across the genome. RP helped us identify a possible mode of inheritance for these new loci.

Aging well is the goal of most Americans. It is clear from data gathered from those reaching the age of 100 and beyond (centenarians) that at least some of the liability for longevity is inherited. We looked at cardiovascular and other functioning in the offspring of centenarians and a control group to determine what, if any, advantages they had. It is well established that body mass index (BMI) plays a significant role in living well to an old age. Our data were self-report data and were incomplete for some study participants. The missing data made traditional models with list-wise deletion difficult. It was important to replicate the well-known BMI association with longevity in order to lend credence to our other reports. Using surrogates in CART, we were able to demonstrate the effect of low BMI on longevity as well as other cardiovascular advantages of the centenarian offspring.

Jason Haukoos MD, Denver Health Medical Center

"CART for Outcome Predictions in Clinical Settings: Emergency Department Triage, Survival Prediction and Prediction of Neurologic Survival"

Dr. Haukoos has significant experience using Classification and Regression Tree (CART) analysis and will discuss its use in biomedical research. Dr. Haukoos will discuss methodological and statistical details of three completed studies, emphasizing specific CART-related modeling techniques. The first study used CART to derive a clinical decision instrument to help triage HIV-infected patients upon presentation to the emergency department. The second study used CART to identify optimal cutoff points for predictors of survival in patients diagnosed with colon and rectal carcinoma in order to categorize patients into low- and high-risk groups. Finally, the third study used CART to derive a clinical decision instrument using patient characteristics available to paramedics to predict meaningful survival following out-of-hospital cardiac arrest.

BIOMEDICAL PRESENTATIONS (continued)

Stuart Gansky, UCSF School of Dentistry

"MARS and Related Techniques in Dental Biomaterials Modeling"

A recent study of mechanical properties of the dentinoenamel junction (DEJ) of the human tooth (Marshall et al, 2001) sought to estimate the width of the interface of two dissimilar materials based on atomic force microscope-derived nanohardness and elastic modulus measurements. Here we study the statistical techniques used for that estimation, including restricted cubic splines, local polynomial regression (loess), adaptive linear basis functions, and parametric change point models. Piecewise linear models are used to simulate data with a plain, rising slope, and plateau with a known horizontal distance (width) between the plain and plateau and varying amounts of error. Five hundred replicates per combination of simulation conditions are used. Statistical techniques to estimate this width are compared to assess bias and efficiency.

Donovan Chin, Biogen

"Improved Predictions in Structure-Based Drug Design Using CART and Bayesian Models"

This talk will describe our use of CART for improving the accuracy of predicting active and non-active compounds using virtual screening and structure-based drug design methods. Virtual screening involves using high performance computations and calculations on large databases of compounds to predict those that will be "successful" against a target of therapeutic interest. We will discuss the following two metrics of success involving CART. First, the ability to predict compounds that will interact most favorably with the 3D structure of the target. Second, the ability to predict compounds that will have desirable drug or bioavailability properties. The background and challenges of the rational drug-design problem, and the lessons learned using CART, will be discussed.

Kenna Mawk, Ciphergen Biosystems, Inc.

"Mining SELDI ProteinChip Data for Biomarkers and Disease Stratification"

Based on patented Surface Enhanced Laser Desorption/Ionization (SELDI) technology, Ciphergen's ProteinChip Systems offer a single, unified platform for a multitude of proteomics research applications. In particular, the use of SELDI technology is expanding rapidly in clinical proteomics to identify proteins that may be useful as biomarkers and to correlate these potential biomarkers with disease states. However, protein profiling presents a significant challenge for statistical analysis, as typical differential protein expression studies via SELDI generate 100's to 1000's of spectra that must be sifted through to find patterns correlating to phenotype or to find individual biomarker candidates. In practice, biomarkers may be expressed as single markers or multiple markers whose patterns of up- and down-regulation may signal disease susceptibility, onset, progression, therapeutic response, drug efficacy or toxicity. We have incorporated CART analysis to simplify analysis of these complex data sets, identify patterns associated with disease and create optimal profiles for patient stratification. This paper will review the scientific basis for the method and demonstrate key results generated in recent studies with medical research collaborators.

Jing Wang, Ph.D., Cengent Therapeutics Inc.

"CART in Drug Discovery: Rules for Making Better Small Molecules"

Application of data mining methods in drug discovery is a highly promising yet almost unexplored field. As one of the first to apply CART in the drug discovery field, we have used CART in developing PTP1B inhibitors as potential diabetes drugs. The first usage was to isolate the major physicochemical properties of compounds that play determinant roles for the kinetic behavior of compounds in the enzyme inhibition. This helped to design competitive inhibitors. The second usage was to discover the structural features of compounds responsible for their inhibitory activities to PTP1B. The derived features are similar to the concept of "pharmacophore models" and can be used in the prediction and design of better compounds.

Brydon Grant, University of Buffalo School of Medicine

"Using MARS for the Prediction of the Apnea-Hypopnea Index"

Obstructive sleep apnea affects 2-4% of the adult population in the USA. Various indices have been developed to assess the presence or absence of sleep apnea from changes of arterial oxygen saturation recorded overnight. This presentation will show how we developed a mathematical model that would combine the best predictive properties of these indices. We obtained overnight oximetry during overnight polysomnography in 224 patients and validated it prospectively in another 292 patients. We developed a single MARS model using the indices of pulse oximetry that has the best predictive properties and an aggregation of 20 MARS models. The aggregated model had greater diagnostic utility than the single MARS model and is currently in clinical use at our medical center.

BIOMEDICAL PRESENTATIONS (continued)

John Warner, Novartis
"Drug Discovery Clinical Trials and Random Forests at Novartis"

This presentation discusses the use of the random forest methodology, a tree based procedure that makes use of bootstrapping and random feature generation (Breiman, 2001), in the mining of a pooled phase II and III Novartis clinical trials database. The goal of this data mining exercise is to construct a predictive model that explains the variation in clinical response to drug associated with patient demographics, medical histories, concomitant medications, laboratory measurements, and/or adverse events. Predictive models of this sort can be used to inform clinical development decision making, address regulatory inquiries, or to generate publications in support of marketing efforts. The main challenge presented by this type of data is the need to screen a large number of variables (1000-5000) for main effects and interactions – a task for which random forests is well suited. I will provide a brief introduction to the industrial context and rationale for using random forests including data set properties, relevant features of the random forest algorithm, and followup analyses performed on the screened predictors. The methodology is illustrated using a disguised Novartis clinical trials database.

The following presenters have chosen to release their presentations at a later date. Their presentations are NOT included in this CD. Their abstracts can be found on page 11.

Wayne Danter, Critical Outcome Technologies, Inc.
"21st Century Drug Discovery Using Hybrid CART, MARS, and Neural Network Models"

Shenghan Lai, Johns Hopkins Bloomberg School of Public Health
"Examples in Epidemiology: Using CART, MARS, and TreeNet"

Debopriya Das, Cold Spring Harbor Laboratory
"Application of MARS to Gene Expression Data: Predictive Models of Gene Regulation"

FINANCIAL SERVICES AND MARKETING PRESENTATIONS

Charles Pollack, Suncorp Metway

"Insurance Premium Increase Optimization: Case Study"

Synopsis: When applying insurance premium change to a group of customers, different groups of customers have differing levels of price elasticity and hence a different point at which the price change is so great that they don't renew their policy. This case study examines a two-stage process involving the use of CART to identify various groups of customers and Logistic Regression to model the elasticity of the different groups. The resulting model then allows appropriate caps on premium increase to be applied to the different customer groups, thereby maximizing customer profitability.

Conclusion: An insurance portfolio is made up of many different groups of customers. These customers have their own decision processes and sensitivities. A one-size-fits-all capping process is hardly an optimal way to achieve the aim of easing price upheaval and optimizing customer renewal. Indeed, capping the level of increases can be a very costly decision. A process that can take account of the sensitivities of various customer groups to reduce this cost, yet is easily explained to management, is a very valuable tool. The use of CART to develop definitions of customer groups free from preconceptions, combined with Logistic Regression to determine the elasticity of each group, can be used to develop such a process. As has been demonstrated in this case study, the financial benefit for the company is significant.

John Trimble, Wells Fargo

"CART/ MARS Risk Assessment of Automobile Loans and Leases"

This presentation describes two ways in which CART and MARS have been used at Wells Fargo Auto Finance Group. The first is a brief description of how MARS was used to quickly recalibrate an existing model of delinquency assessment, resulting in substantial cost savings. The second is a work in progress. It describes how CART and MARS have been used thus far to understand the underlying attributes of prepayers of automobile loans.

David McCloskey, Pathfinder Solutions Ptd.

"Predicting Customer Behavior Trends Over Space and Time"

Modelling for customer acquisition can involve sparse data conditions, limiting the predictive power of the models. Techniques to enrich data for use in CART and MARS models are detailed, highlighting how CART can be used to overcome some obstacles to data enrichment found with other analytical techniques. Particular focus is applied to the use of spatially based information for data enrichment.

Modelling for customer churn, in comparison to modeling for acquisition can involve use of detailed transactional data, captured over a period of months or years. A method of incorporating time series analysis into CART modeling to predict churn is discussed.

Larry Lai, DIRECTV, Inc.

"Variable Derivation and Selection For Customer Churn Models"

In a CRM environment, it is not uncommon to have a huge data warehouse containing all customer touches through different contact channels- mail, phone call, e-mail and website, inbound or outbound. Data content could be either origination at the time of registration such as credit risk, dealer channels, geographic, lifestyle, psychographic and demographic or longitudinal over life such as customer consumption, payment and contact. Prior to a CART modeling analysis, it can be more productive to conduct variable derivation and selection within a subcategory of attributes with similar context first, e.g. within billing or customer contact category as opposed to the entire database. This presentation describes the benefits of some techniques for conducting "partial" variable derivation and selection before engaging in a CART analysis and illustrates with a real case of customer churn prediction model.

FINANCIAL SERVICES AND MARKETING PRESENTATIONS (continued)

Edward Malthouse, Northwestern University

"High-Level Marketing Strategy: Deciding Whether to Customize Product Offerings and their Promotion and Whether to Reward "Best" Customers with Perks"

"Data mining" has been successful in optimizing existing processes, tasks, tactics, etc. in a wide variety of fields, including marketing. Optimizing such tasks can result in very large cost savings and/or incremental profit. But data mining has not been used in making high-level (marketing) strategy or policy decisions within an organization. This talk discusses whether data mining can and should play a role in such decisions. Data mining fundamentally studies how to make accurate predictions. I give two examples where understanding predictive accuracy is crucial to making high-level marketing decisions: deciding whether a firm should (1) customize product offerings and their promotion and (2) reward alleged "best customers" with perks/superior benefits. I give examples using Salford System's software of the two problems with real data from a credit card, software, retail catalog, and educational service companies, and a not-for-profit organization. This talk ultimately argues that there are many additional applications of data mining in marketing and challenges the audience to expand data mining's influence in an organization by contributing to the solution of strategic problems.

Louise Francis, Actuarial Analytics

"Insurance Fraud Detection: MARS vs. Neural Networks"

A recently developed data mining technique, Multivariate Adaptive Regression Splines (**MARS**) has been hailed by some as a viable competitor to neural networks that does not suffer from some of the limitations of neural networks. Like neural networks, it is effective when analyzing complex structures which are commonly found in data, such as nonlinearities and interactions. However, unlike neural networks, **MARS** is not a "black box", but produces models that are explainable to management. This paper will introduce **MARS** by showing its similarity to an already well-understood statistical technique: linear regression. It will illustrate **MARS** by applying it to insurance fraud data and will compare its performance to that of neural networks.

Jon Farrar, Union Bank of California

"Union Bank's Use of CART to Identify Customer Attrition and Screen Application Fraud"

This presentation will illustrate how the bank is using CART to help create applied solutions to various types of real world business challenges.

The presentation overviews three (and time permitting, five) main challenges:

1. Small Business Application Fraud
2. Customer Attrition
3. Creating Decision Rule scripts to explain model development results
   a. Non-standard models such as Neural Networks and TreeNet
   b. Facilitate understanding for Internal Audit, Compliance and/or Regulators

(Time Permitting)

4. CART as a variable selection aid
5. CART as a Customer association or clustering aid

FINANCIAL SERVICES AND MARKETING PRESENTATIONS (continued)

Beng-Hai Chea, Citibank, N. A.

"Committee of Decision Trees Solution for Personal Bankruptcy Prediction"
****A modified version of this paper will be presented at "Salford Data Mining 2005".***

Abstract: Characterization of individual behaviours by the process of data mining has been key tool for commercial banks to manage credit losses, for a long time. Many techniques are available and used by either in-house research teams or Commercial Score Developers. However all these techniques are not very effective when it comes to predicting personal bankruptcies. In current unstable environment when technology is progressing in a phenomenal speed, structural unemployment is becoming a major issue in many countries. Coupled with this other economic turmoil, globalisation is pushing many individuals towards bankruptcy who were considered highly credit worthy only a few years ago. Another problem with bankruptcy is the suddenness of its occurrence making the prediction extremely difficult. Hence long term credit history, which has been the backbone of successful data mining, is often useless in this case. All of these has created a strong need of advanced/ non-standard models for predicting personal bankruptcy risk so that it is possible to avoid high credit losses by the consumer banks. Quite a number of papers/articles/journals/books are available on the topic of corporate bankruptcies and the usage of Artificial Neural Networks, Discriminant Analysis and Logit Analysis for such predictions, however there are only a handful of articles written on personal bankruptcies and the usage of committee of decision trees for the prediction purpose. This presentation discusses in details how the committee of decision trees can successfully predict personal bankruptcies where the standard scoring techniques or the decision trees fail. The solution has been successfully validated with real in-house data.

*****The following presenter did not want his presentation released. The presentation is NOT included in this CD. The abstract can be found on page 11.***

David Goldsmith, MDT Advisers

"Combined Time Series and Cross Sectional CART Modeling for Common Stock Selection"

This presentation shows how CART can be used to cluster common stocks based on predicted relative performance. Although standard CART is a useful tool for this problem, it does not really address the statistical subtleties inherent in modeling combined time-series and cross-sectional (TSCS) data. The good news, though, is that the split rules and forecasting methods in standard CART can be readily extended to more accurately model TSCS data. The presentation will describe a way of doing this for my TSCS problem, and suggests that the same general approach can be used for other statistically complicated problems.

ENVIRONMENTALLY ORIENTED PRESENTATIONS

Cecil Dharmasri, Syngenta Crop Protection, Inc.

"The Importance of CART and MARS in Environmental Fate and Risk Assessment for Pesticides"

Syngenta is a world-leading agribusiness committed to sustainable agriculture through innovative research and technology. The company is a leader in crop protection and ranks third in the world in the high-value commercial seeds market. We are evaluating the use of CART and MARS in environmental studies to understand the underlying complex processes governing environmental fate of pesticides. By using CART and MARS we hope to uncover important relationships among environmental fate of pesticides, weather, soil, and agricultural practices. CART will be used to identify the key environmental factors and their interactive effects on the fate of pesticides. MARS will be used to develop models to predict the environmental fate of pesticides for the greatly varying soil and weather conditions throughout the world. These analyses will lead to understanding of the spatial and temporal patterns in the environmental monitoring databases. With this information, we would be able to refine the environmental risk assessment for pesticides, both protecting the environment while more effectively using pesticides. Both CART and MARS give easily understandable insights to the environmental data and reveal hidden relationships in environmental databases and will serve as complementary tools to mechanistic predictive modeling. This presentation will show the potential of CART and MARS in improving the understanding of the complex spatial and temporal environmental processes. Some examples of CART and MARS applications will be presented with details on underlying environmental databases and challenging tasks of data mining.

Belle Hudischewskyj, ICF Consulting/Systems Applications International, Inc.

"Application of CART for Air Quality Forecasting"

This presentation provides an overview of a recent study carried out for the Mid-Atlantic Regional Air Management Association (MARAMA). The primary objective of the study was to develop and deliver documented and tested methods for forecasting particulate matter concentration of less than 2.5 microns in diameter (known as $PM_{2.5}$) for several cities in the MARAMA region. These methods were then applied in the development of a real-time forecasting tool for $PM_{2.5}$. Exposure to $PM_{2.5}$ has been linked with adverse health affects, and increasing numbers of areas are initiating forecasting programs that provide warnings to the public of days for which high levels of these particulates are expected. Many current air quality-forecasting algorithms rely on simple regression techniques. While such techniques may provide an analytic description for some of the data dependence, they are not likely to give much physical insight. For example, a result of "3 times temperature plus 0.5 times the relative humidity equals $PM_{2.5}$ concentration" is not particularly intuitive, nor can it be reasonably checked against known physical principles. In this study, a CART-based technique was applied and used in the development of a forecasting tool. The CART-based technique, unlike the regression technique, provides physical insight as well as results that can be reviewed relative to known physical principles and then used to develop/enhance conceptual models. Thus, using this technique, we establish a physically meaningful basis for $PM_{2.5}$ forecasting. We applied CART to data representative of nine cities within the MARAMA region (Philadelphia, PA, Baltimore MD, Washington DC, Richmond VA, Roanoke VA, Bristol VA, Wilmington DE, Charlotte NC, and Newark NJ.) for the period 2000 – 2002. (Longer periods are generally preferable, however for many of the cities, historical records of monitored $PM_{2.5}$ were not available prior to 2000.) Separate CART trees were created for each of the cities. These CART trees placed days with different values of the classification variable into bins representing different ranges of $PM_{2.5}$ concentrations. In the application of CART, the dependent variable was the 24-hour average $PM_{2.5}$ concentration. Other air quality and meteorological variables that would be readily available from other sources at the time of the forecast, comprised the set of independent variables. Once the CART trees were created, they were transformed into user-friendly forecasting algorithms/tools that would allow air quality forecasters to input observed and predicted input variables via PC-based forms. These forms would then automatically transfer the data to the tool that would then place a future day into a classification bin. Based on the category of the bin, the forecaster would have the information necessary to issue a forecast on the anticipated $PM_{2.5}$ concentration for the day in question. The presentation illustrates the use of the forecasting tool via graphical displays and summarizes some of the key relationships identified between the dependent and independent variables.

ENVIRONMENTALLY ORIENTED PRESENTATIONS (continued)

Christopher Martius, University of Bonn

"Assessing the Sustainability of Agroforestry Systems Using CART to Model Non-linear Relationships"

The sustainability of agroforestry systems was assessed by looking at termite assemblages in the litter layer and topsoil of these systems. Termites are important ecosystem engineers and decomposers that act in nutrient cycling and soil formation. We studied four sites in central Amazonia, a primary rain forest, a 13-year old secondary forest, and two sites of an agroforestry system that consisted of four planted tree species and upcoming secondary vegetation. Termite diversity (13 genera), abundance (1.1C103 ind. m-2) and biomass (0.7 g m-2) were significantly higher in the primary forest than in the secondary forest and plantation systems (8-9 taxa; abundance 0.2-0.5 C103 ind. m-2; biomass 0.1-0.3 g m-2). Litter termites and smaller species were primarily those reduced. We assessed several environmental variables as potential factors responsible for these differences in termite biomass (dead wood volume, litter stocks, canopy closure, litter and soil site average temperatures, and several microclimate variables immediately (3-30 days) before the termite sampling). For any of these variables, the linear regressions with litter, soil or total termite biomass yielded either low correlation coefficients, low sensitivity, or both. However, non-linear relationships between termite biomass and environmental variables were found by applying a “classification and regression tree” (CART) approach. Depending on the model setup, different CART trees either effectively separated high from low termite biomass (litter, soil, or total) at a threshold of canopy closure of 87.5% (separating primary forest from the other sites), or at a threshold of an average pre-sampling soil temperature of 27/C during three days before sampling. We conclude that: 1) environmental variables that control soil macrofauna may often be non-linear and rather be characterized by reaction to thresholds of the environmental variables that characterize survival limits; 2) in agroforestry sites, termite diversity and biomass may be reduced because temperature extremes are more likely to occur; 3) agroforestry systems should maximize canopy closure to preserve the beneficial sub-surface decomposer communities.

Falk Heuttmann, University of Alaska

"Predictive and Spatial Modeling Applications for Wildlife Research and Conservation Management: An Overview Linking CART and MARS with GIS"

Predictive spatial modeling is a steeply growing multidisciplinary science. Applications are manifold and reach from plain distribution questions over investigations of abundance and populations up to Population Viability Analysis (PVA), global change modeling and others. This presentation tries to give a brief overview about applications, benefits and unresolved questions when using advanced modeling algorithms such as CART and MARS in the context of wildlife research and conservation management. Drawing from modeling applications and examples world-wide it is shown how these progressive algorithms can be linked with a Geographic Information System (GIS), and how model inference and model predictions can be derived. Conceptual comparisons with other and more traditional algorithms are done as well. The importance of model evaluation is discussed, as well as the use of ‘presence only’ data for such modeling projects, e.g. based on Natural History Collections (NHC), field survey and telemetry data. An outlook is given on the role that predictive spatial modeling will probably play in the future, and how it can be used for decision support on a global scale, e.g. when combining it with interactive web-based approaches.

PRESENTATIONS NOT INCLUDED IN THIS CD
***The following presenters have chosen to release their presentations at a later date.**

Wayne Danter, Critical Outcome Technologies, Inc.
"21st Century Drug Discovery Using Hybrid CART, MARS, and Neural Network Models"

The traditional drug development process is lengthy, expensive, inefficient and unnecessarily risky. It costs on average more than 500,000,000.00 USD, takes at least a decade and 5000-10000 compounds must be screened in order to take 1 successful drug through to regulatory approval. The pre-clinical phase of this process relies heavily on in vitro (i.e. cellular systems) and in vivo (i.e. animal) testing.

Powerful computers and sophisticated modeling software like CART®, MARS™ and Neural Networks combined with a growing understanding of the complex and highly nonlinear relationship between molecular structure and biological activity have started a paradigm shift away from traditional methods towards in silico pre-clinical drug profiling.

Hybrid non-linear machine learning systems that combine CART®, MARS™ and Artificial Neural Network models of the molecular structure-biological activity relationship can accurately simulate specific in vitro and in vivo biological methods that have traditionally been used to predict efficacy, drug-like properties, pharmacokinetic profile, metabolic profile and acute toxicity.

The main impacts of in silico drug profiling are (1) a reduction in the pre-clinical phase from the current 4-6 years to ~18 months, (2) a reduction in the attrition rates for screened molecules and (3) a dramatic reduction in the overall cost of developing a successful drug.

Debopriya Das, Cold Spring Harbor Laboratory
"Application of MARS to Gene Expression Data: Predictive Models of Gene Regulation"

Gene transcription is an essential process in the survival of living organisms. Regulation of gene transcription is complex, especially in eukaryotes. It is executed by a large number of transcription factors (TFs) which bind to the specific DNA elements in the promoter regions of the genes. It is inherently combinatorial in nature. Thus, in many cases, multiple transcription factors bind to the promoter region of a gene and synergistically activate or repress transcription. Current computational models for transcriptional regulation do not take such synergistic effects into account. Older approaches to this problem have been restricted to clustering genes by their expression profiles followed by a search for a common set of DNA motifs in the promoter regions of the genes in a given cluster. This approach is limited by the fact that many genes do not tend to cluster. Recent attempts have tried to overcome this limitation by fitting a linear model between the frequencies of binding motifs and the gene expression levels. However, such models do not take into account the synergistic effects described above. Besides, since in such an approach, the binding affinities of various TFs are approximated by the motif frequencies, there are additional complex dependencies which are not accounted for. MARS (Multivariate Adaptive Regression Splines) provides a natural framework to quantify such cooperative effects along with the linear effects on gene expression. MARS breaks down the feature (motif) space into separate domains of smooth variability and obtains the best fit in each such domain: the final prediction is an additive contribution from each such domain. Thus, along with the nonlinearities and interactive effects, it can capture additional complexities in the underlying data. We apply MARS to the microarray expression data and show that inclusion of synergistic pairs leads to an improvement of fit over the noninteracting model in 70% of the cases, with an average improvement of 46% in such cases. Significant motifs and motif pairs at various stages of transcription are appropriately predicted.

Shenghan Lai, Johns Hopkins Bloomberg School of Public Health
"Examples in Epidemiology: Using CART, MARS, and TreeNet"

MARS, CART and TreeNet/MART are exceptionally powerful techniques in analyzing massive epidemiologic data. In this presentation, several examples from our epidemiologic research are used to illustrate the usefulness of these data mining techniques. The first example was to use MARS to explore the association between regional heart function and coronary calcification. This example demonstrated that without MARS analysis, the conventional approach failed to identify the association. The second example was to use CART to classify the study participants. This example showed that CART is more powerful than the conventional logistic regression analysis. The third example was to use MART to explore the association between vitamin E and the development of myocardial infarction. Again, this example suggested that without MART, the relationship may never be able to be identified.

David Goldsmith, MDT Advisers
"Combined Time Series and Cross Sectional CART Modeling for Common Stock Selection"

This presentation shows how CART can be used to cluster common stocks based on predicted relative performance. Although standard CART is a useful tool for this problem, it does not really address the statistical subtleties inherent in modeling combined time-series and cross-sectional (TSCS) data. The good news, though, is that the split rules and forecasting methods in standard CART can be readily extended to more accurately model TSCS data. The presentation will describe a way of doing this for my TSCS problem, and suggests that the same general approach can be used for other statistically complicated problem.