Final Report
2003–04 Program on
Data Mining and Machine Learning (DMML)

June 13, 2006

1 Summary

The high-level goals of the SAMSI DMML program were to advance significantly the understanding fundamental statistical and computational issues in data mining, machine learning and large data sets, to articulate future research needs for DMML, especially from the perspective of the statistical sciences, and to catalyze the formation of collaborations among statistical, mathematical and computer scientists to pursue the research agenda.

By almost every measure, the program was a strong success. Among high points:

- A deeper understanding of the points of connection between data mining and statistical theory and methodology.
- Effective analyses of a large, extremely complex testbed database provided by General Motors (GM), and affiliate of NISS—and therefore also SAMSI, strengthening SAMSI's industrial connections.
- A strong and continuing collaboration in the area of metabolomics, involving chemists, computer scientists and statistical scientists, leading to publications, a proposal and several program participants' being asked to serve of the scientific advisory board of Metabolon.
- A range of specific progress on issues ranging from false discovery rates to overcompleteness to support vector machines.
- Profound effects on the career development of participants. One postdoctoral, who took leave from a faculty position to participate in the program, is being reviewed for tenure this year, and feels that SAMSI strengthened his case significantly.

  Another, more senior, participant stated: "Much to my wonderment, I realized that yes, this is it, the place I'd dreamed of. Full of energetic people with insights, imagination, and a hard-charging curiosity about whatever the world has to offer. [...] This is the best intellectual/research environment I've ever had the chance to enjoy. It seemed to me that everyone is keen to bring in the future, teach their fellow travelers and learn in turn, and enjoy the development of new ideas."

There are, as in any SAMSI program, ways in which the program could have achieved broader impact. The most significant missed opportunity was that only as the program developed did

realization crystallize of the need for applied mathematics engagement, in keeping with the SAMSI vision. There was a relative dearth of truly complex testbed databases. The two database analyzed in detail were too complex to serve as testbeds for evaluation theory and methodology developed in the program. (To some degree, however, the expectation that multi-gigabyte databases would serve as testbeds was simply too optimistic.) Only the analysis of the one GM testbed database represented impact of the program on the practice of data mining, but other impacts will come with time.

2 Program Review

2.1 Program Leadership

The Scientific Committee for the program consisted of David Banks (Duke; co-chair), Mary Ellen Bock (Purdue; NAC liaison), Jerome Friedman (Stanford), Alan F. Karr (NISS; co-chair and Directorate liaison), David Madigan (Rutgers), William DuMouchel (AT&T), Warren Sarle (SAS Institute).

2.2 Program Goals

The principal objectives of the DMML program were to:

- Advance significantly understanding of fundamental statistical and computational issues in DMML;
- Articulate future research needs for DMML, especially from the perspective of the statistical sciences;
- Catalyze the formation of collaborations among statistical, mathematical and computer scientists to pursue the research agenda;
- Employ databases provided by NISS Affiliates as testbeds to evaluate existing and new DMML tools, as well as furnish useful analyses to the owners of the testbeds;
- Engender community interest and engagement in the program, through workshops, research visits and the project Web site (www.samsi.info/200304/dmml/dmml-home.html).

3 Working Groups

Scientific activities of the program occurred primarily in five working groups, which had distinct but overlapping foci. The "Large p, Small n Inference" and "Theory and Methods" working groups have collaborated closely and often met together. Each group met at least weekly throughout the year, and in addition there were weekly meetings of the program co-chairs (Banks and Karr) with the working group leaders.

At the start of the year, in connection with preparation of the programs "Goals and Outcomes Document," each working group was asked to identify one or more outcomes that it would consider to be "stunning successes," as well as a detailed research agenda.

3.1 Bioinformatics

This working group was led by Stanley Young, Assistant Director of NISS. Other participants were Chris Beecher (Metabolon), Atina Brooks (graduate student, North Carolina State University (NCSU)), Jun Feng (postdoc, NISS), Jacqueline Hughes-Oliver (NCSU), Gerardo Hurtado (SAS Institute), Xiaodong Lin (postdoc, SAMSI and NISS), Andrew Nobel (University of North Carolina at Chapel Hill (UNC)), Katja Remlinger (graduate student, NCSU), Susan Simmons (University of North Carolina at Wilmington), Alexander Tropsha (UNC), Young Truong (UNC) and Michiel van Rhee (ICAGEN).

The group adopted as an organizing principle the drug discovery pipeline—target, identification, assay development, high throughput screening, secondary endpoint prediction, lead optimization, clinical trials and epidemiology.

Goals. Two "stunning successes" were identified:

1. The lead scientist at a local startup has mass spectroscopy data metabolomics data on 1000 small molecules, with $n \ll d$ with many values measured at approximately 0. Inference needs include prediction of disease state. Because scientifically metabolomics lies beyond proteomics (which in turn lies beyond microarrays), this is a major opportunity for early injection of statistics into a new and important area.
2. Expansion of high throughput screening (HTS) data analysis into detection and exploitation of synergistic compounds. A collection of n compounds has $\sim n^2$ pairs of compounds. Searching for and finding bioactive pairs of compounds is a great opportunity.

Research Agenda. Specific objectives were to:

- Assemble model data sets;
- Collect, review and disseminate key software, algorithms, techniques and papers;
- Identify important, approachable statistical problems;
- Sketch papers to write.

Milestones included securing data sets, securing small company collaborators and submitting papers.

Achievements:

- A new method of determining the key binding features of compounds to a protein. A provisional patent has been filed, a paper is in preparation, and Jun Feng will present the results at the annual American Chemical Society meeting.

- Studies of cross validation when $n \ll p$ and there are twin observations. In chemistry data sets, there are often very similar compounds—"twins"—and these can cause usual methods of cross validation, e.g., leave-one-out, to be misleading. The working group is critiquing a *PNAS* paper, studying theory papers and conducting simulations to study this situation.
- Assembling a number of data sets used for benchmarking prediction methods where the data sets are unbalanced—there are few active observations and many inactive observations.
- A proposal (unsuccessful) to the NIH for research on data analysis in metabolomics was submitted in March, 2004 to the National Institutes of Health (NIH). Participants were Banks, Hughes-Oliver, Young, Lin, House, Truong, Adele Cutler (Utah State) and Susan Simmons (UNC Wilmington).

  As a result of the collaboration leading to this proposal, Stanley Young, Young Truong, David Banks, and Jackie Hughes-Oliver are serving on the scientific advisory board of Metabolon, Inc.

Publications:

- D. Banks, J. Woo, D. Burwen, P. Perucci, M. Braun, and R. Ball (2005). Comparing data mining methods on the VAERS database. *Phamacoepidemiology and Drug Safety* **14** 601–609.
- C. Beecher, A. Cutler, L. House, X. Lin, Y. Truong, and S. S. Young (2004). Learning a complex metabolomic dataset using random forests and support vector machines. *Proceedings of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- J. Feng, A. Sanil, and S. S. Young (2006). PharmID: Pharmacophore identification using Gibbs sampling. *J. Chem. Inf. Model.* **46(3)** 1352–1359.
- D. M. Hawkins, R. D. Wolfinger, L. Liu, and S. S. Young (2003). Exploring blood spectra for signs of ovarian cancer. *Chance* **16** 19–23.
- L. Liu, D. M. Hawkins, S. Ghose, and S. S. Young (2004). Robust singular value decomposition analysis of microarray data. *Proc. Nat. Acad. Sci.* **100** 13167–13172.
- S. J. Simmons, X. Lin, C. Beecher, Y. Truong, and S. S. Young (2004). Active and passive learning to explore a complex metabolism data set. *Classification, Cluster Analysis, and Data Mining*, 447–457. Springer-Verlag, Berlin.
- Y. Yang and G. A. Rempala (2005). A Note on multiple tests for gene expression data. *Journal of Statistical Planning and Inference* (to appear, subject to revisions).
- S. S. Young and N. Ge (2004). Design of diversity and focused combinatorial libraries in drug discovery." *Current Opinions in Drug Design and Development* **7(3)** 318–24.
- S. S. Young, M. Wang and F. Gu (2003). Design of diverse and focused combinatorial libraries using an alternating algorithm. *J. Chem. Info. Comp. Sci.* **43** 1916–1921.

3.2 GM Data Analyses

This working group was led by Alan Karr. Other participants were David Banks, Ashish Sanil (NISS), Peter Westfall (Texas Tech), Jen-hwa Chu (graduate student, Duke) and more than a dozen researchers, analysts and managers from GM.

Because of the special relationship between GM and NISS/SAMSI, planned analysis of three testbed databases was structured as a cross-cutting activity. For a variety of reasons, only one database, containing vehicle sales data, was analyzed in detail. Warranty data lacked sufficient detail, and a set of manufacturing plant monitoring data never materialized.

Goals. The most important goal was to produce actionable scientific insight for GM, derived from a combination of exploratory and simple analyses of the testbed databases, application of existing DMML tools and use of DMML tools developed by the program, some of which respond directly to needs raised by the GM data.

Research Agenda. As noted above, only one of an anticipated three testbed databases was analyzed in detail: *demand sensing* data concerning vehicle sales. Discussions with GM led to a single initial question:

- What factors—such as vehicle characteristics (type, options, ...) and geography—affect time_to_turn, the time between when a dealer receives a vehicle and when it is sold to a customer?

Other sub- and related questions were raised as well, for example, whether time_to_turn differs between dealer-order and customer-ordered vehicles. (It does.)

Achievements. The scale and complexity[1] of the demand sensing data were much greater than anticipated, so to a significant extent this database served less as a testbed for sophisticated DMML tools than as a means of demonstrating the effectiveness of exploratory analyses. Specific achievements were:

- Tools for managing and manipulating the data, which included relational database management systems (RDBMSs), statistical packages and customized scripts.
- Detailed study of differences of time_to_turn between dealer-ordered and customer-ordered vehicles.
- Maps showing the median time_to_turn by state and vehicle brand, which showed that brand effects on time_to_turn dominate geographical effects.
- An insightful volume $\times$ time_to_turn classification of brands, which has been adopted by GM.
- Regression analyses of time_to_turn at the merchandising model level, identifying options that affect time_to_turn. Not only the scale but in some cases even the sign of the effect are merchandising model-specific.

[1] More than 2.5 million vehicles and more than 1200 option codes in more than 500,000 combinations.

These achievements occurred despite significant masking of the data by GM, which was necessary in order to make the data available to SAMSI.

Publication:

- A. F. Karr and A. P. Sanil (2005). Analysis of demand sensing data. Delivered to GM on December 21, 2004.

Discussions are in progress with GM to create a version of this report suitable for the open literature.

3.3 Large p, Small n Inference

This working group was led by Bertrand Clarke (British Columbia), SAMSI–University Fellow. Other participants were David Banks, Prem Goel, M. J. Bayarri (Valencia), Dongchu Sun (Missouri), Merlise Clyde (Duke), Andrew Nobel (UNC), Ashish Sanil (NISS), Feng Liang (Duke), Yuguo Chen (Duke), Ernest Fokoué (postdoc, SAMSI), Xiaodong Lin (postdoc, SAMSI and NISS), Murali Haran (NISS), Jesus Palomo (Madrid), Fei Liu (graduate student, Duke), Jen-hwa Chu (graduate student, Duke) and Eric Vance (graduate student, Duke).

This working group focused on inference in the "large p, small n" setting in which the number of dimensions in the data exceeds, perhaps by orders of magnitude, the sample size. As noted above, it and the Theory and Methods working group often met and worked together.

Goals.[2] An identified stunning success was to construct a matrix of techniques and measures of performance. The top row of the matrix would list various techniques such as clustering, classification, regression, survival analysis, model averaging and multivariate methods in general. (Within each of these categories further distinctions could be nested. For instance, classification contains random forests, support vector machines (SVM), neural nets and distance-weighted discrimination.) Down the left-hand column would be a variety of measures of performance such as prediction error, interpretability, computational efficiency, scalability and so forth. Entries in the matrix would be derived from extensive theoretical or computational comparisons of diverse existing methods.

Research Agenda. Specific objectives included:

- Investigating model uncertainty through a general bias-variance decomposition;
- Further study of model averaging;
- Development of new methods.

Achievements. Substantial progress has been made on a number of issues:

- Effective sample size and/or effective parameter size work by Clarke and Lin and by Clarke and Ao Yuan (Howard University), leading to one paper.

[2] These are the same goals articulated by the "Theory and Methods" working group.

- Using prediction optimality for function approximation/model uncertainty in a machine learning context. This work by Clarke and Fokoué and by Clarke and Steven Wang (York University) led to two papers and a monograph chapter.
- Regression on statistics in a "large p, small n" context by Clarke and Chu.

This working group also conducted an ongoing seminar on statistical issues in DMML, which included presentations by Murali Haran on spatial statistics, by Andrew Nobel on clustering, by Ernest Fokoué on model uncertainty, by Susie Bayarri on hypothesis testing, by Merlise Clyde on overcompleteness and by Feng Liang on overcompleteness.

Publications:

- E. Fokoué (2004). Parsimonious function representation and optimal predictive model selection. SAMSI technical report number 2004–19. Submitted to *Canadian Journal of Statistics*.
- E. Fokoué (2004). "Stochastic determination of the intrinsic structure in Bayesian factor analysis" SAMSI technical report number 2004–17.
- E. Fokoué (2004). Sparsity through prevalence estimation. Submitted to *J. Machine Learning Res.*
- E. Fokoué and B. Clarke (2004). Optimal model list selection for prediction. SAMSI technical report number 2004–20.
- X. Lin, J. Pittman and B. Clarke (2004). Bayesian sample size and effective parameter size. Submitted to *Journal of the American Statistical Association*. SAMSI technical report number 2004–21.

3.4 Support Vector Machines

This working group was led by Marc Genton (NCSU), Faculty Fellow. Other participants were Jeongyun Ahn (graduate student, UNC), Ernest Fokoué (postdoc, SAMSI), Prem Goel (Ohio State), Gerardo Hurtado (SAS Institute), Xiaodong Lin, Peng Liu (graduate student, NCSU), (postdoc, SAMSI and NISS), J. S. Marron (SAMSI and UNC), Cheolwoo Park (SAMSI), Dongchu Sun (Missouri), Young Truong (UNC) and Helen Zhang (NCSU).

This was perhaps the most technical of the working groups, with a highly focused research agenda. However, there were significant interactions with the Bioinformatics working group.

Goals. Development of workable, interpretable multi-category SVM was identified as a particular notable success. It was achieved in part.

Research Agenda. This working group identified a series of "clusters" of interests, each with a local leader and a set of research goals:

- Multi-category SVM, to investigate SVM methods for multi-category classification problems, including ordered classes, for example, survival times in biomedical applications.

- Kernel choice for SVM, addressing such issues as the importance of the choice of the kernel for performance of SVM methods, whether identity kernels suffice for some applications and the gain from using compactly supported kernels.
- Feature selection SVM, possibly also addressing missing values problems (in high dimensions) for SVM, mixed data and interpretability of SVM methods.
- Space-time data mining, to investigate the use of SVM methods for space-time data.

Achievements. Progress was made on number of fronts:

Bayesian SVM: A rigorous statistical justification for the Relevance Vector Machine (RVM). Many interesting and promising ideas for such a characterization have arisen, which will be made more concrete in a paper entitled "On some statistical properties of the relevance vector machine and related methods."

A paper proposing a hierarchical structure for the RVM is being finalized. The main result is that the extended prior structure will make it possible to obtain a unique solution. Mathematical expressions for the posteriors of interest have been derived and written up, and the next step is to code the scheme and test it on various examples.

A new method had been developed for finding a sparse representation of an approximating function—a fully Bayesian treatment of kernel expansion and basis expansion using a hierarchical structure. Computationally, the method combines a birth-and-death process with a Gibbs sampling updating move to estimate the number of prevalent vectors or basis elements as well as those vectors or basis elements themselves.

The members of cluster also helped in the preprocessing of the MonoAmineOxidase (MAO) dataset. MAO data were analyzed using both traditional SVM and the RVM.

Feature selection: a new regularization method for variable selection in SVM, which replaces the lasso-type L^1 penalty by a nonconcave penalty called SCAD (smoothly clipped absolute deviation). Experimental studies using the gene expression data set and the metabolite data set, show that SCAD-SVM works very well in terms of classification error and selecting the important features. Two cross validation methods to select the tuning parameter are were investigated, as are generalizations of the circle of ideas to nonlinear SVM.

Multi-category SVM: Investigation SVM methods for multicategory classification problems, including improvements of proximal SVM. Applications were made to the Reuters data set and ordered classes, for example, survival times in biomedical applications.

Kernel selection: Expansion and testing of a compactly supported kernel approach. A working paper on the topic can be obtained from Genton.

Space-time data: Identification of a data mining and machine learning strategy for a space-time database, furnished by the National Center for Atmospheric Research (NCAR), containing 150,000 hourly observations and two responses, one of which is categorical and the other

continuous, mandating use of at least two kernels. There is also severe autocorrelation among the variables. Approaches under investigation include multi-stage and sub-sampling.

Publications:

- E. Fokoué, P. Goel, and D. Sun (2004). A new hierarchical prior structure for the relevance vector machine.
- H. H. Zhang, J. Ann, X. Lin and C. Park[3] (2005). Gene selection via learning with concave penalty. *Bioinfomatics* **22(1)** 88–95.
- H. Zhang, J. Ahn, X. Lin, and C. Park (2005). Variable selection for SVM using shrinkage methods. In B. Clarke, ed., *Principles and Theory of Data Mining*.
- H. H. Zhang, M. Genton and P. Liu. (2004). Compactly supported radial basis kernel. Under revision for *Journal of Machine Learning Research*.

3.5 Theory and Methods

This working group was led by David Banks and Prem Goel (Ohio State), Senior Fellow. Other participants were Bertrand Clarke, Chris Beecher (Metabolon), M. J. Bayarri (Valencia), Dongchu Sun (Missouri), Merlise Clyde (Duke), Andrew Nobel (UNC), Ashish Sanil (NISS), Feng Liang (Duke), Susan Simmons (University of North Carolina at Wilmington), Greg Rempala (University of Louisville), Yuguo Chen (Duke), Ernest Fokoué (postdoc, SAMSI), Xiaodong Lin (postdoc, SAMSI and NISS), Murali Haran (NISS), Jesus Palomo (Madrid), Fei Liu (graduate student, Duke), Jen-hwa Chu (graduate student, Duke), Eric Vance (graduate student, Duke), Leanna House (graduate student, Duke), and Balaji Krishnapuram (graduate student, Duke).

The focus of this working group was to break the disconnect between existing DMML tools and rigorous understanding of their properties from a statistical perspective.

Goals.[4] An identified stunning success was be to construct a matrix of techniques and measures of performance. The top row of the matrix would list various techniques such as clustering, classification, regression, survival analysis, model averaging and multivariate methods in general. Within each of these categories further distinctions could be nested. For instance, classification contains random forests, SVM, neural nets and distance-weighted discrimination. Down the left-hand column would be a variety of measures of performance such as prediction error, interpretability, computational efficiency, scalability and so forth. Entries in the matrix would be derived from extensive theoretical or computational comparisons of diverse existing methods.

Research Agenda. Identified research emphases were:

- Unlabeled samples for training classification algorithms, to understand statistically how to take subtle advantage of implicit information in the unlabeled cases to enhance the performance of trained classifying rules.

[3] A postdoc in the 2003–04 Internet Traffic program, which illustrates synergies among SAMSI programs.

[4] These are the same goals as for the "Large p, Small n" working group.

- Overcompleteness: many data mining methods that work well use much more than a minimal orthogonal basis of functions in doing their fits and predictions. To statisticians this seems to create a need for regularization or shrinkage, generates multiple testing problems, and may prevent the discovery of interpretable structures. The group will examine these issues to find out why gross expansion of the set of fitting functions seems to work.
- X-raying black boxes, using statistical methods to understand and interpret the "black boxes" built by computer scientists.
- Computer experiments, which would be designed simulation experiment to compare methods, probably in the context of some specific class of problems (e.g., microarray cluster analysis).

Achievements. Principal scientific achievements to date are:

- Text mining to infer Bureau of Labor Statistics (BLS) occupational categories for Census long-form answers, which was completed in April, 2004.
- Research on robustness in data mining, including new ideas in overcompleteness that may improve the kernel trick via a Bayesian technique. A batch of smaller ideas and insights have been developed that may grow into something more substantial over time, including the "twin problem" in cross-validation, better than training performance with semi-labeled data, and use of false-discovery rate methods to control effect of multiple decisions in data mining.

Publications:

- D. Banks (2004). A dimension reduction technique for local linear regression. In *Classification, Cluster Analysis, and Data Mining*. Springer–Verlag, Berlin.
- D. Banks and L. House (2004). Robust multidimensional scaling. *Proceedings in Computational Statistics 2004*, 251–260. Physica–Verlag, Berlin.
- D. Banks, L. House, P. Arabie, F. R. McMorris, and W. Gaul, eds. (2004). *Classification, Clustering, and Data Mining*. Springer–Verlag, Heidelberg.
- D. Banks and F. Liang (2004). Review of *The Elements of Statistical Learning* by T. Hastie, R. Tibshirani, and J. Friedman. *Journal of Classification* **21(10)** 155–157.
- D. Banks, J. Woo, D. Burwen, P. Perucci, and R. Ball (2005). Comparison of four methods of data mining in the vaccine adverse event reporting system. *Phamacoepidemiology*, to appear.
- M. J. Bayarri, J. Berger, G. Garcia-Donato, F. Liu, J. Palomo, R. Paulo, J. Sacks, D. Walsh, J. Cafeo, and R. Parthasarathy (2006). Computer model validation with function output. Submitted to *Annals of Statistics*.
- M. J. Bayarri and G. García-Donato (2006). Extending conventional priors for testing general hypothesis in linear models. *Biometrika* (to appear).

- B. Clarke, ed. (2006). *Principles and Theory of Data Mining*. Submitted to ASA–SIAM Series on Statistics and Applied Probability.
- R. Derrig and G. Rempala (2005). Claims in the presence of suspicion of fraud and build-up. A statistical analysis of the settlement negotiation process for automobile bodily injury liability. Submitted to *Journal of Risk and Insurance*.
- L. House and D. Banks (2004). Cherry-picking as a robustness tool. In D. Banks, L. House, P. Arabie, F. R. McMorris, and W. Gaul, eds., *Classification, Cluster Analysis, and Data Mining*, 197–208. Springer–Verlag, Berlin.
- D. Jeske and Regina Liu (2004). Mining massive text data and developing tracking statistics. In D. Banks, L. House, P. Arabie, F. R. McMorris, and W. Gaul, eds., *Classification, Cluster Analysis, and Data Mining*, 495–510. Springer-Verlag, Berlin.
- X. Lin and Y. Zhu (2004). Degenerate expectation-maximization algorithm for local dimension reduction. In D. Banks, L. House, P. Arabie, F. R. McMorris, and W. Gaul, eds., *Classification, Cluster Analysis, and Data Mining*, 259–268. Springer-Verlag, Berlin.
- G. Rempala and J. Wesolowski (2005). Asymptotics For products of independent sums with an application to Wishart determinants. *Statistics and Probability Letters* **74(2)** 129–138.
- S. J. Simmons, X. Lin, C. Beecher, Y. Truong, and S. S. Young (2004). Active and passive learning to explore a complex metabolism data set. In D. Banks, L. House, P. Arabie, F. R. McMorris, and W. Gaul, eds., *Classification, Cluster Analysis, and Data Mining*, 447–457. Springer-Verlag, Berlin.

4 Workshops and Courses

Seven workshops were held as part of the DMML program, together with one semester-long course and one week-long technology transfer course.

4.1 Tutorial and Kickoff Workshop

The DMML program was initiated by three back-to-back-to-back events on September 6–10, 2003, which served to focus the scientific agenda of the program, as well as highlight the statistical importance of work by non-statisticians in such areas as support vector machines.

Tutorials, which introduced important topics to both experienced and new researchers:

- *Large p, Small n Inference*, by David Banks of Duke University
- *Support Vector Machines*, by J. S. Marron of SAMSI and the University of North Carolina at Chapel Hill.

Kickoff Workshop, which in addition to ten invited presentations listed in the attached supporting material, featured a number of innovations designed to maximize participation of all attendees. These included:

- Birds-of-a-Feather Sessions reflecting workshop and participant interests, which served as precursors of the Working Groups.
- Poster Sales Talks, allowing each poster presenter to introduce his or her topic.
- Poster Session, at the NISS/SAMSI building.
- Second Chance Seminar, at which anyone could talk, which focused on curricular issues involving data mining and machine learning.
- New Researchers Session, at which seven students, postdoctoral fellows and new faculty members presented their research. One senior participant said that this session "restored my faith in the future of the field."

Working Group Meetings, on September 10, at SAMSI, which the working groups were able to draw on the ideas on other Kickoff Workshop participants to formulate initial research agendas.

4.2 Mid-Year Workshops

Three mid-year workshops were held, tied to the working groups:

Support Vector Machines: January 28, 2004

Large p, Small n Inference/Theory and Methods: February 4, 2004

Bioinformatics: February 11, 2004.

The purposes were to:

- Assess progress over the fall semester, as well as problems encountered;
- Set the high-level research agenda for the spring semester; and
- Provide an opportunity for the statistical sciences community to learn about progress to date, provide feedback and become engaged in spring activities.

Participants included DMML visitors, postdocs, students and local faculty and other researchers who are part of the working groups, as well as 2–3 invited outside speakers at each workshop. The atmosphere was informal, highly participatory and intense. Program details appear in the attached supporting material.

4.3 Closing Workshop

The closing workshop for the program was held on May 17–18, 2004. It served two principal functions:

- To present both results and a research generated by the DMML program to the statistical sciences, applied mathematics and computer science communities.
- To formulate follow-on activities for the program, and specifically to engage attendees who did not participate deeply in the program in these activities.

There were approximately 50 participants. The program appears in the attached supporting material.

4.4 Undergraduate Workshops

Two undergraduate workshops entitled "Data Mining: Handling the Flood of Data" were held, on November 14–15, 2003 (30 attendees) and February 13–14, 2004. The purposes were to introduce undergraduates to DMML using adaptive, interactive demonstrations. The workshops feature multiple problem contexts, including bioinformatics (drug discovery), software engineering (data from instrumented software) and the GM sales data. Both underlying concepts, some of which are quite simple despite the extreme computational demands and current research frontiers, such as privacy preserving data mining, were covered. Program details appear in the attached supporting material.

4.5 Semester-Long Course

Feng Liang and David Banks taught a semester-long advanced graduate course in DMML to 43 students. The students came from all three area universities (Duke, North Carolina State and UNC) and were pursuing Ph.D. work in statistics, computer science, and electrical engineering. There were also regular auditors from SAS and Glaxo-SmithKline.

Half the course was drawn from the later chapters of Hastie, Tibshirani, and Friedman's book *The Elements of Statistical Learning*. Specifically, material covered included support vector machines, the kernel trick, Vapnik-Chervonenkis theory, multidimensional scaling and SOMs. The first part of the course was based on lecture notes that reviewed smoothers, nonparametric regression, the Curse of Dimensionality, projection pursuit and related algorithms, neural nets, MARS, CART, and dimension reduction.

4.6 Technology Transfer Short Course

The goals of the DMML technology transfer short course, which was held on July 25–29, 2005, were to:

- Provide a survey of the theoretical basis for modern data mining

- Give participants hands-on experience with data mining software
- Convey insights and strategies for data mining practice.

Principal instructor for the course was David L. Banks, Professor of the Practice of Statistics and Decision Sciences at Duke University, and co-leader of the DMML program.

The structure of the short course was three hours of lecture each morning. Each afternoon started with a 90-minute computer lab that goes over an application using real data and relevant software, followed by a 90-minute lecture by a guest speaker.

The guest speakers were:

Jack Liu (GlaxoSmithKline): Visualization and Data Mining for Microarrays

J. S. Marron (UNC at Chapel Hill): Issues with High Dimensional, Low Sample Size Data

Feng Liang (Duke): Model Complexity and Regularization

Merlise Clyde (Duke): Bayesian Model Averaging.

Course contents were:

1. Background and Overview: Nonparametric Regression, Cross-Validation, the Bootstrap
2. Key Ideas and Methods: Smoothing, Bias-Variance Tradeoff
3. Search and Variable Selection: Experimental Design, Gray Codes, Fitness
4. Nonparametric Regression: Heuristics on Eight Methods
5. Comparing Methods: Designing Experiments in Data Mining
6. Local Dimension: How to Pick Problems Wisely
7. Classification: Boosting, Random Forests, Support Vector Machines
8. Cluster Analysis: Hierarchical, k-Means, and Mixture Models; SOM
9. Issues with Bases: Hilbert Space, Shrinkage, Overcompleteness
10. Wavelets: Introduction, Construction, Examples
11. Structure Extraction: Regression and Multidimensional Scaling
12. Vapnik-Cervonenkis Classes and PAC Bounds

5 Personnel

5.1 Postdoctorals

Two SAMSI postdoctorals were appointed for this program:

Ernest Fokoué (Ph.D., Glasgow; on leave from Ohio State) participated principally in the SVM and Theory and Methods working groups, leading work on Bayesian SVM. He gave multiple presentations that summarized work from the Machine Learning Conference in August, 2003, on overcompleteness and on variational methods.

Xiaodong Lin (Ph.D., Purdue; assuming a faculty position at the University of Cincinnati in September, 2004) played a key role in the metabolomic analyses, became proficient with Breiman and Cutler's random forest code, Hawkins' singular value decomposition techniques, and several flavors of SVM. He led work on feature selection for SVM. In addition, Lin maintained the group's web site, and gave various presentations to the group on topics such as dimension reduction and the kernel method.[5]

Other postdoctorals from NISS and elsewhere were regular participants in DMML activities:

Jun Feng, NISS postdoc (Ph.D., Medicinal Chemistry, UNC) has participated in many aspects of the Bioinformatics working group.

Murali Haran, NISS postdoc (Ph.D., Minnesota; assuming a faculty position at Penn State University in September, 2004) participated in the Large p, Small n and Theory and Methods working groups.

Jesus Palomo (Madrid) participated in the Large p, Small n and Theory and Methods working groups, and gave presentations on the false discovery rate and other multiple comparison methods in the context of data mining and structure discovery.

5.2 Research Visitors

Research visitors to SAMSI for the DMML program, with affiliations, dates and roles, were as follows:

M. J. Bayarri, Valencia: Multiple times throughout the year, to participate in the Large p, Small n and Theory and Methods working groups.

Sudip Bose, George Washington University: March 10-11, 2004, to discuss general issues in data mining.

Song Chen, Iowa State University: February 4, 2004, to attend the mid-year workshop.

Hugh Chipman, University of Waterloo: February 2–4, 2004, to present work of his on tree-structured inference at the February 4 mid-year workshop, and to discuss plans for a similar program to be sponsored by the Canadian National Program on Complex Data Structures in October of 2004.

James Cox, SAS: February 4, 2004, to present work on text mining at the mid-year workshop.

[5] Lin was supported jointly by NISS (25%) and SAMSI (75%); his work at NISS dealt with data confidentiality.

Adele Cutler, Utah State University: February 8–14, 2004, to initiate collaboration with SAMSI personnel on random forests, to participate in the metabolomics project, and to attend the February 11 mid-year workshop.

Jerome Friedman, Stanford University: October 6–8, 2003, for general discussions and to present the SAMSI Distinguished Lecture on October 6.

Prem Goel, Ohio State University: September–December 2003, to participate in the Theory and Methods working group.

Giles Hooker, Stanford University: February 4, 2004, to present research on functional data analysis at February 4 mid-year workshop.

Karen Kafadar, University of Colorado at Denver: September 6–10, 2003, to participate in Kick-off Workshop, and December 15–19, 2003, to discuss overcompleteness.

Ravi Khatree, Oakland University: April 1–30, 2004, to participate in text mining on Census occupational data.

Liza Levina, University of Michigan: December 8–14, 2003, to speak on the surprising success of the naive Bayes classifier.

Regina Liu, Rutgers University: February 4, 2004, for general discussions and to speak on text mining in airline safety reports.

Yvonne Martin, Abbott Labs: February 11, 2004, for research discussions and to participate in the Bioinformatics Working Group mid-year workshop.

Thomas Mitchell, Carnegie Mellon University: March 3, 2004, for general discussions and to speak on fMRI analysis.

Kerby Shedden, University of Michigan: February 10–12, 2004, for research discussions and to participate in the Bioinformatics Working Group mid-year workshop

Dongchu Sun, University of Missouri: October 2003 and February–March, 2004, to participate in research on MCMC for data mining.

Jiayang Sun, Case Western Reserve University: October 12–15, 2003, for research discussions and presentation on text mining and multiple comparisons.

William Welch, University of British Columbia: multiple visits in connection with the Bioinformatics working group.

Tong Zhang, IBM: January 27–29, 2004, for research discussions and to participate in the January 28 mid-year SVM workshop.

Ji Zhu, University of Michigan: January 27–29, 2004, for research discussions and to participate in the January 28 mid-year SVM workshop.

6 Follow-On Activities

Proposals. The Theory and Methods and Bioinformatics working groups have produced one proposals, in the area of data analysis in metabolomics, that was submitted in March, 2004 to the NIH. It was not funded. Other proposals are under development.

A proposal from Clarke to the Natural Sciences and Engineering Research Council (Canada) for follow-on research was funded in full.

Discussions with GM regarding a NISS-led follow-on project are under way.

Dissertations. Jen-hwa Chu (Duke) is writing a dissertation on a data mining topic started with Bertrand Clarke. Leanna House (Duke) is writing a dissertation on proteomics with Merlise Clyde that grew from work on metabolomics problems.

Monograph. A monograph, tentatively entitled *Data Mining: Principles and Methods* on data mining is being written that pulls together research from all four working groups. It is planned to be submitted to the *ASA/SIAM Series on Statistics and Applied Probability*. David Banks will be its editor. Approximately ten individual papers are anticipated:

1. Basics, covering the "curse of dimensionality," cross-validation, bootstrapping and bias-variance tradeoff, all from a regression standpoint
2. Search strategies: model, list, combinatorial, geometric
3. Smoothing: short review of kernels, splines, . . ., leading to matrix formulation
4. Classical techniques: CART, ACE, MARS, PPR, neural nets
5. New techniques: SVM, random forests, MART, boosting
6. Classification
7. Cluster analysis
8. Support vector machines
9. Variable selection for SVM
10. Heuristics for large p, small n
11. Variational methods.

Electronic Frontier Foundation (EFF) Panel. The EFF has contracted with NISS to convene an expert panel to review the technical feasibility of prospective data mining of public records, taking into account effects of

Data integration: the combining, often imperfectly, of multiple, "related" databases, often assembled by different organizations for different purposes; and

Data quality: the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions.

These are problems, together with data confidentiality,[6] in which National Institute of Statistical Sciences (NISS) has been deeply engaged for the past several years. The panel, therefore, leverages the complementary strengths of NISS and SAMSI. David Banks is co-chairing that panel, together with Stephen Fienberg of Carnegie Mellon University (CMU).

Conference Sessions. Research arising from the data mining and machine learning program has been has been presented at a number of conferences and international meetings, often in sessions specifically focusing on the SAMSI data mining and machine learning program:

- MD-2003 (A SAS conference in 2003): David Banks
- 2004 Quality and Productivity Research Conference: Susan Simmons, Leanna House, Jacqueline Hughes–Oliver, Stanley Young, Ashish Sanil
- 2004 Interface Conference (Baltimore): Ernest Fokoué, David Banks, Ashish Sanil
- International Society for Bayesian Analysis (ISBA): Bertrand Clarke, Feng Liang, Merlise Clyde, Susie Bayarri
- 2004 Spring Research Conference on Statistics in Industry and Technology (Raleigh) : Feng Liang, Jen-hwa Chu, David Banks
- International Federation of Classification Societies (Chicago, July 15–18, 2004): Helen Zhang, Xiaodong Lin, Leanna House, Stanley Young, Jacqueline Hughes–Oliver
- 2004 SIAM Annual Meeting (Portland; July 12–16, 2004): David Banks
- 2004 Joint Statistics Meetings (Toronto; August 8–12, 2004): David Banks; also, a topics contributed session organized by Murali Haran with Bertrand Clarke, Ernest Fokoué, Leanna House, and Katja Remlinger.
- COMPSTAT'04 (Prague; August 23–27, 2004): David Banks
- Fields Institute Workshop (NPCDS—Canada; Toronto, October 28–30, 2004): David Banks, Merlise Clyde, Adele Cutler
- Army Conference on Applied Statistics (Atlanta, October 20–22; 2004): David Banks
- International Conference on the Future of Statistical Theory (Hyderabad, India; December 27, 2004–January 2, 2005): David Banks, Prem Goel, Murali Haran.

[6] The need to balance *disclosure risk* (of data subject identities and attribute values) against *data utility* to multiple constituencies, including federal agencies, researchers and the public.

Workshops. A metabolomics workshop at NISS is planned for the summer of 2005.

Education and Outreach. A short course on data mining was presented by David Banks JSM 2004, and again at the Hyderabad conference in December 2004.

SAMSI Technology Transfer Course. Banks is developing a one-week technology transfer course on DMML, to be offered at SAMSI in the summer of 2005.

7 Outcomes and Measures of Success

A complete evaluation of the DMML program will be included in the 2005 SAMSI Annual Report.

7.1 Program Level

Outcomes at the program level were:

- Significant research accomplishments by the working groups, leading to papers submitted during or shortly after the program year.
- Scientific insight resulting from analysis of testbed databases provided by GM and Metabolon.
- Formation of new collaborations, leading to proposals and research in following years.
- Extremely positive career impact on participants, especially postdoctoral researchers.
- Strong community interest in the program, leading to engagement in the form of research visits and workshop participation.
- An extensive report detailing the conclusions and recommendations of the program, to be published in the SAMSI subseries of the *ASA–SIAM Series on Statistics and Applied Probability*.

Each of these has clear measures of success, based on either quantifiable information (e.g., numbers of papers and proposals produced, numbers of visitor) or participant self-assessment (e.g., feedback from database providers, follow-up with postdoctoral fellows). Not all of these measures, however, operate within the one-year time frame of the program, and so post-program follow-up will be necessary.

7.2 Working Groups

As described above, each working group was asked to identify

- One or more outcomes that it would consider to be "stunning successes." Not all did so, but each has addressed high-level goals.
- A detailed research agenda that both focuses energy and defines measures of success.

Every working group made significant progress on its research objectives, leading to at least one and in most cases multiple publications and continuing collaborations.

Supporting Material

Kickoff Workshop Presentations. Invited presentations at the September 7–9, 2003 Kickoff Workshop were:

Leo Breiman, University of California, Berkeley: Similarities and Differences between Statistics, Machine Learning and Data Mining

Leo Breiman, University of California, Berkeley: Statistical Tools for the Sciences

Di Cook, Iowa State University: Using Graphics in Exploratory Data Analysis and Data Mining—An Application of Supervised Classification in Olive Oil Quality

William DuMouchel, AT&T Labs Research: Postmarketing Drug Adverse Event Surveillance and the Innocent Bystander Effect

Michael I. Jordan, University of California, Berkeley: Convex Optimization and Variational Inference Algorithms—Alternatives to MCMC for Large-scale Statistical Models

David Madigan, Rutgers University: Statistical Methods for Text Mining Data

Robert McCulloch, University of Chicago: Bayesian Additive Regression Trees

Jeff Schneider, Carnegie Mellon University: Mining in Anti-Terrorism Applications

K. P. Unnikrishnan, General Motors Research and Development: Temporal Data Mining—Novel Algorithms and Their Applications

Peter Westfall, Texas Tech University: Using Proc MULTEST of SAS/STAT for Data Mining

Mid-Year Workshops. Programs for the three mid-year workshops follow below.

Support Vector Machines: January 28, 2004

9:00–9:05 AM	Welcome to SAMSI Alan Karr, NISS and SAMSI
9:05–10:05	Piecewise Linear SVM Paths Ji Zhu, University of Michigan
10:15-10:45	Bayesian Analysis of SVM and SVM-like Techniques: The Present and Some Ideas for the Future Ernest Fokoué, SAMSI
10:45–11:00	Bayesian SVM Discussion Leader: Ernest Fokoué
11:15–11:45	Compactly Supported Kernels Helen Zhang, NCSU
11:45–12:00 N	Kernel Selection Discussion Leader: Marc Genton, NCSU
1:00–1:30 PM	Variable Selection for SVM using SCAD Penalty Xiaodong Lin, SAMSI
1:30–1:45	Feature Selection Discussion Leader: Cheolwoo Park, SAMSI
2:00-2:30	Mining Space-time Data Peng Liu, NCSU
2:30–2:45	Space-Time Discussion Leader: Marc Genton
3:00–4:00	Statistical Models for Binary and Multi-category Large Margin Methods Tong Zhang, IBM
4:15–5:00	Panel Discussion: *Open Problems and Future Directions* Atina Brooks, Helen Zhang, Tong Zhang and Ji Zhu

Theory and Methods: February 4, 2004

9:00–9:30 AM	Introduction to SAMSI and the Data Mining Year
9:30–10:00	Overcompleteness Feng Liang, Duke, and Fei Liu, Duke
10:00–10:30	Large p, Small n Bertrand Clarke, SAMSI
10:45–11:15	Multiple Testing David Banks, Duke
11:15–11:45	GM Data Jen-hwa Chu, Duke
1:15–2:00 PM	Invited Talk Giles Hooker, Stanford
2:00-2:45	Invited Talk Hugh Chipman, Waterloo
3:00–3:45	Invited talk Jim Cox, SAS
3:45–4:30	Mining Massive Text Data and Developing Tracking Statistics Regina Liu, Rutgers
4:30-5:00	Summary

Bioinformatics: February 11, 2004

9:00–9:05 AM	Welcome
9:05–9:20	Bioinformatics Year; Virtual Screening Stan Young, NISS
9:20–10:20	Invited Talk Yvonne Martin, Abbott Laboratories
10:20–10:35	Discussion
10:50–11:20	Invited Talk Alex Tropsha, UNC
11:20–11:30	Discussion
11:30–12:00 N	New Method for Pharmacophore Mapping Jun Feng, NISS
12:00–12:10	Discussion
1:10–2:10 PM	Towards Interpretability of Classifiers for Virtual Screening Will Welch, University of British Columbia
2:10–2:25	Discussion
2:25–2:55	Structure/Activity Analysis of Chemosensitivity Variation based on Bond Arrangements Kerby Shedden, University of Michigan
2:55–3:05	Discussion
3:20–3:40	SVM applied to MAO Dataset Atina Brooks, NCSU Scott Oloff, UNC
3:40–3:50	Discussion
3:50–4:35	Applications of virtual screening as used by ∗ The Hereditary Disease Foundation: Jun Feng ∗ LDDN: Ke Zhang, NCSU
4:35–4:45	Discussion
4:45–5:15	Directed discussion and summary of the day's events Yvonne Martin, Alex Tropsha, Jacqueline Hughes–Oliver

Closing Workshop Program

Monday, May 17, 2004

9:00 AM	Welcome and Introductions
9:15	*Working Group I: Bioinformatics* Summary: Stanley Young, NISS Technical Highlight: Jacqueline Hughes-Oliver, North Carolina State University Problem List: Young Truong, University of North Carolina at Chapel Hill
10:30	Discussion
10:45	Break
11:00	*Working Group II: Large n, small p* Summary: Bertrand Clarke, SAMSI Technical Highlight Ernest Fokoué, SAMSI Problem List Feng Liang, Duke University
12:15 PM	Discussion
12:30	Lunch
1:30	*Working Group III: Support Vector Machines* Summary: Marc Genton, North Carolina State University Technical Highlight Helen Zhang, North Carolina State University Problem List Ernest Fokoué, SAMSI
2:45	Discussion
3:00	Break
3:30	*Working Group IV: Theory and Methods* Summary: David Banks, Duke University and SAMSI Technical Highlight Merlise Clyde, Duke University Problem List Xiaodong Lin, SAMSI and NISS
4:45	Discussion
5:00	“Open Mike” Discussion
6:30	Reception

Tuesday, May 18, 2004

9:00 AM	Panel Discussion I: Working Group Leaders
10:00	Break
10:15	Graduate Student Presentations Atina Brooks, North Carolina State University Jen-hwa Chu, Duke University Leanna House, Duke University Fei Liu, Duke University Peng Liu, North Carolina State University
11:30	Panel Discussion II: Selected “Outsiders” Hugh Chipman, University of Waterloo William DuMouchel, AT&T Labs Research Douglas Hawkins, University of Minnesota David Madigan, Rutgers University
12:30 PM	Adjourn

Undergraduate Workshop Program. The program for the November 14–15, 2003 and February 13–15, 2004 undergraduate workshops "Data Mining: Handling the Flood of Data" follows below.

Day 1

10:00–10:15 AM	Welcome and Introductions H. T. Banks, SAMSI and NCSU Alan Karr, NISS and SAMSI
10:15–10:45	About SAMSI and NISS Alan Karr
10:45–12:30	Introduction to Data Mining Alan Karr
12:30 N–1:15 PM	Lunch
1:15–2:45	Mining Software Engineering Data Ashish Sanil, Research Statistician, NISS
2:45–3:00	Break
3:00–4:30	Mining Pharmaceutical Data Stanley Young, Assistant Director for Bioinformatics, NISS
4:30	Adjourn for the Day

Day Two

10:00–11:00 AM	Starting to Mine "Real" Data Alan Karr
11:00–12:00 N	Related Problems: Data Confidentiality and Data Quality Alan Karr
12:00 N	Adjourn