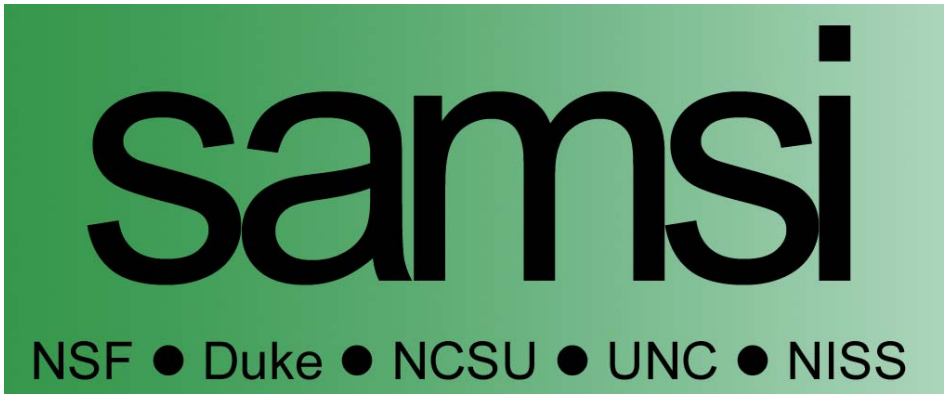



# Annual Scientific Report

## 2003-2004

***May 1, 2004***

# SAMSI Annual Scientific Report for 2003-2004

This report is a version of the SAMSI Annual Report to the National Science Foundation, with sensitive financial data and personal information removed. It covers the period of SAMSI activities from July 1, 2003 – June 30, 2004. Past and future activities of SAMSI are also discussed.

## 0. Executive Summary

The Executive Summary contains

A. Outline of SAMSI Activities and Initiatives for Year 2 and the future
B. Financial Summary
C. Directorate's Summary of Challenges and Responses
D. Synopsis of Developments in Research and Education
E. Evaluation by the SAMSI Governing Board.

## A. Outline of Activities and Initiatives

### 1. Second Year Programs and Activities

*Focused Study Programs*

- Network Modeling for the Internet (Fall 2003-Spring 2004)
  - Meeting on Measurement and Modeling (9/19/03)
  - Workshop on Internet Tomography and Sensor Networks (10/12/03-10/15/03)
  - Workshop on Congestion Control and Heavy Traffic Modeling (10/31/03-11/1/03)
  - Closing Workshop (6/24/04-6/26/04)
- Multiscale Model Development and Control Design (Spring 2004)
  - Tutorials and Opening workshop (1/17/04-1/20/04)
  - Workshop on Multiscale Challenges in Soft Matter Materials (2/15/04-2/17/04)
  - Workshop on Fluctuations and Continuum Equations for Granular Flow (4/16/04-4/17/04)

*Synthesis Program*

- Data Mining and Machine Learning (Fall 2003-Spring 2004)
  - Tutorials and Opening workshop (9/6/03-9/9/03)
  - Midterm Workday on Support Vector Machines (1/28/04)
  - Midterm Workday on Theory and Methods & Large p, Small n Inference (2/4/04)
  - Midterm Workday on Bioinformatics (2/11/04)
  - Closing workshop (5/17/04-5/18/04)

*Education and Outreach*

- Industrial Mathematical and Statistical Modeling Workshop for Graduate Students (7/21/03-7/30/03)
- Two-Day Workshops for Undergraduates (11/14/03-11/15/03 and 2/13/04-2/14/04)
- Undergraduate Interdisciplinary Workshop (5/31/04-6/4/04)
- Industrial Mathematical and Statistical Modeling Workshop for Graduate Students (7/26/04-8/3/04)
- Graduate Courses at SAMSI
    - Data Mining and Machine Learning, Fall 2003
    - Data Statistical Analysis and Modeling of Internet Traffic Data, Fall 2003
    - Long-range Dependence and Heavy Tails, Fall 2003
    - Multiscale Model Development and Control Design, Spring 2004

*Distinguished Lecture Series*

- Margaret Wright, "Direct Search Methods: The Sound and the Fuss" (10/7/03)
- Jerome H. Friedman, "Importance Sampling: An Alternative View of Ensemble Learning" (11/4/03)
- Jonathan Chapman, "A hierarcy of models for type-II superconductors" (3/2/04)
- Thomas G. Kurtz, "Particle representations of continuum models" (4/6/04)

*Program Planning / Hot Topics Workshops*

- High Dimensional Inference and Random Matrices (2/29/04-3/1/04)
- Mathematical and Statistical Research for National Security (4/1/04-4/2/04)
- Design and Analysis of Computer Experiments for Complex Systems (7/13/04-7/17/04)

## 2. Third Year Program Schedule

*Focused Study Programs*

- Computational Biology of Infectious Disease (Fall 2004-Spring 2005)
    - Opening Tutorials and Workshop (9/18/04-9/22/04)
    - Mid-Program Focused Workshop (TBA)
    - Transition Workshop and Symposium (5/22/05-5/24/05)
- Latent Variable Models in the Social Sciences (Fall 2004-Spring 2005)
    - Tutorials and Opening workshop (9/11/04-9/15/04)
    - Mid-Program Focused Workshop (TBA)
    - Closing Workshop (5/19/05-5/21/05)
- Data Assimilation for Geophysical Systems (Spring 2005)
    - Tutorials and Opening Workshop (1/23/05-1/26/05)
    - Issues, Challenges & Interdisciplinary Perspectives (joint at IPAM, 2/22/05-2/26/05)
    - Summer School - "Fusing Models with Data: From Theory to Practice to Theory" (joint at NCAR, 6/12/05-6/23/05)

*Education and Outreach*

- Two 2-Day Workshops for Undergraduates will be held during the academic year
- An Undergraduate Interdisciplinary Workshop will be held in early June 2005

- The Industrial Mathematical and Statistical Modeling Workshop for Graduate Students will be held in late July 2005
- Graduate Courses at SAMSI
    - Kepler/Lindsey, Fall 2004
    - Alun Lloyd, Fall 2004
    - Lloyd Edwards, Fall 2004
    - Goldstein, Spring 2005
    - Data Assimilation, Spring 2005

*Tentative Programs for 2005-2006*

- Mathematical and Statistical Research for National Security (Fall 2005 and Spring 2006)
- High Dimensional Inference and Random Matrices (Fall 2005)
- Mathematical and Statistical Finance (Fall 2005)
- Astrostatistics (Spring 2006)

# 3. Developments and Initiatives

*Second-Year Developments*

- Planning workshops were instituted as a way to provide focus for future programs, and greatly improved the program development process.
- The NISS University Affiliates program was converted to a joint NISS/SAMSI University Affiliates program, and the affiliates have had a significant effect on programs (see, e.g., the Internet and Data Mining/Machine Learning programs)
- The website is being redesigned, and will include an interactive component to allow programs and workshops to have considerable advance planning.
- Two-day outreach workshops were initiated, for undergraduates from around the country, successfully promoting the SAMSI vision.
- Wireless networking was installed at SAMSI.
- New Researcher Fellowships were established, to be given to new researchers (typically non-tenured or recently tenured), visiting for a semester or a year.
- A collaboration with the American Institute of Mathematics was instituted in the area of Random Matrices; a SAMSI planning program was held at AIM and program interactions will continue.
- A biweekly postdoctoral seminar was instituted, and proved very successful, both in allowing postdocs to better share their scientific developments and interests and to aid in the development of their presentation skills.

*Planned Third-Year Developments*

- Another staff person will be added, in part to deal with the more intensive evaluation schemes that were approved, and in part to systematize program development activities.
- Additional collaborations with other institutes are being planned, to enhance the overall impact of mathematics and statistics; these initiatives include
    - activities with the National Center for Atmospheric Research, relating to the Data Assimilation program, including joint postdoctoral appointments and a planned joint summer graduate educational program;

- a joint workshop with the Centre de Recherches Mathématiques in financial mathematics;
- a joint workshop with the Institute of Pure and Applied Mathematics in data assimilation;
- a variety of coordinated activities with the Canadian National Program on Complex Data Structures, including
    - the upcoming workshop on Design and Analysis of Computer Experiments for Complex Systems, serving as a possible lead-in to a SAMSI program in the area;
    - a co-sponsored Data Mining workshop in Toronto, serving as a significant outlet for dissemination of DMML program results;
    - a potential new researcher meeting, as part of the Social Sciences program, to involve Canadian and U.S. new researchers in the area.

# C. Directorate's Summary of Challenges and Responses

SAMSI has been successful in achieving its goals: the scientific programs have been of high caliber, and have lead to significant new and ongoing research collaborations between, statistics, applied mathematics and disciplinary sciences; there has been significant human resource development, through the postdoctoral and graduate programs and through involvement of senior researchers in new interdisciplinary areas; and many students across the country have been shown the SAMSI vision through educational outreach programs and courses. We feel that these successes are amply demonstrated throughout the report, and will here confine discussion to the challenges that arose in Year 2 and the Directorate's response to these challenges.

*Program Initiation:* Most of the programs conducted during the first two years had been part of the initial SAMSI grant proposal, and hence had local individuals as leaders or co-leaders. This year our focus shifted to the creation of programs to be driven, in large part, by outside scientists. Few programs come 'ready made' with leaders attached. Rather, it is a process of working with key individuals over time, to craft a program of key scientific interest and which they are willing to lead. We are finding that exploratory workshops can be helpful in the process of program creation, and that collaborating with other organizations (e.g., AIM, NCAR, IPAM, and NPCDS, as mentioned earlier) is also valuable in this process.

*Program Development:* As mentioned above, the initial SASMI programs were directed by individuals who had been heavily involved in the creation of SAMSI, and so the process of developing the programs (recruiting participants and postdocs, forming working groups, planning workshops, etc.) was well-understood by the program leaders. The upcoming programs, however, will be lead by individuals with no or minimal previous connection to SAMSI, and they do not know how SAMSI 'works.' Documents were prepared outlining the process of program and workshop development, but we came to realize that these are not sufficient guidance. We are thus instituting other mechanisms to provide more hands-on guidance, including having directorate and NAC liaisons on each program committee; instituting regular meetings or conference calls between the program leaders and

directorate; and hiring another staff person, part of whose responsibility is ensuring that program development keeps to schedule.

*Program Operation and Evaluation:* We are continually adapting program operations to reflect our experiences in running programs. For instance, we observed in Year 1 that standard formats for workshops were not optimal because SAMSI workshops typically involve very diverse communities and have significant planning roles. In Year 2, workshops were accordingly restructured to improve mixing and planning, including having early break-out groups – to start the process of interaction and discussion from the beginning; introducing poster advertisement sessions – wherein presenters at the poster session are each allowed 2 minutes to advertise for their poster; and having 5 minute madness sessions, in which each participant has 5 minutes to discuss their interests and thoughts for the program. A variety of types of structured planning sessions have also been tried.

All workshop participants are asked to submit an evaluation of their experience, and postdocs have been involved in an extensive process of evaluation of their experience. We ensure that responses from postdocs are obtained, but have not had a large response rate from workshop participants. Additionally, feedback from other program participants has been sought, but the response has been irregular – good in some programs and bad in others. As part of the SAMSI Evaluation Plan, approved last November, such evaluations will become more institutionalized. Indeed, the new staff member mentioned above will have a key role in regularizing the evaluations.

As a part of this evaluation, we will do follow-ups to track the impact on research and people of our programs. The only program that ended early enough to make such follow-up feasible at this time was the program on *Inverse Problem Methodology in Complex Stochastic Models.* See this report (Appendix A) for the results of this evaluation.

*Human Resources:* We have generally been successful in achieving the goal of making two-year postdoctoral appointments, with one year at SAMSI and another at a local institution, but this has not proved to be always possible. The principle behind this original goal, however, was to ensure that a postdoctoral fellow would not need to devote significant time to a job search during their postdoctoral experience. We have found that this principle can be satisfied in a variety of other ways, such as having the fellow accept a job, but with a one-year delay to be at SAMSI, or by having joint long-term postdoctoral appointments with non-local institutions (e.g., NCAR).

During the first year of operation, SAMSI was extremely successful in achieving diversity goals, by sustained – but standard – efforts. In the second year, diversity goals were achieved in the education and outreach programs, but were not attained in the scientific programs. There was still a sustained effort in the scientific programs, which did result in significant representation of underrepresented groups, but not at the level we strive to achieve. We are thus alerted to the fact that additional focus is needed to address the problem.

SAMSI Graduate Fellows have the rare opportunity to be immersed in a combined statistics/mathematics/disciplinary interface, and the educational experience is of great benefit to them. SAMSI research programs also require work from graduate students. Finding the right balance between education and research for students, and finding the

students who can most successfully manage both, is a more difficult problem than we had initially realized. Its resolution will require clarification of the role and responsibilities of SAMSI Graduate Fellows, and more extensive discussion with mentors and advisors.

# D. Synopsis of Developments in Research and Education

In later parts of the report, the extensive developments in research and education that have occurred under SAMSI research programs are discussed in detail. To give a flavor of these developments, we highlight some of their findings here.

## 1. Research

a) STOCHASTIC COMPUTATION*:* The SAMSI program on Stochastic Computation of 2002-2003 focused on synthesis and developmental research in four inter-related areas in which realistic applied models are mathematically complex; involve very many uncertain parameters to be estimated; in which the form and specification of the mathematical model itself is uncertain so that scientists need tools for searching over vast spaces of candidate models; and in which data is, relative to model and parameter dimensions, relatively sparse. Some key innovations and results of StoCom include the following.

In the area of *large-scale variable selection and regression model search* – exemplified by problems such as environmental risk assessment – StoCom introduced novel methods for rapid exploration of huge model spaces with efficient simulation algorithms that can substantially improve regression model search when faced with many candidate predictors, and also substantially improve prediction performance as a result.

In problems of *discrete data analysis* using contingency tables subject to problems of high-dimension and missing data – such as arise in applications in areas as diverse as government data-base confidentiality and security, and in human population genetics – StoCom research produced significant advances in computational methods, as well as defining interconnections between branches of core mathematical/algebraic research and more applied statistical research.

In the area of *large-scale graphical models* – such as arise in applications in areas as diverse as genome-scale gene expression studies in modern computational biology, and in large-scale market research studies – StoCom led to substantial advances in the capacity to explore very high-dimensional spaces of graphs, extending theory and computational tools to develop radically more effective than earlier available. This component also explored and generated studies of the approach in exploratory analysis of large-scale genomic data via cluster computing, with resulting software tools provided to the research community.

In the area of *financial modeling*, StoCom brought together statistical and mathematical researchers to define new approaches to stochastic computation in some of the most challenging financial volatility and pricing models, and defined new research directions from this collaboration in studies of volatility at multiple scales.

b) INVERSE PROBLEMS: The SAMSI Inverse Problems Program of 2002-2003 has had a tangible impact in merging ideas from applied mathematics and statistics to develop new techniques to use noisy, incomplete data to estimate parameters in complex HIV models.

When combined with control theory techniques, the resulting models suggest how improved therapies – suboptimal controlled structured treatment interruptions – may be effective in permitting patients to stop therapy for long periods of time without suffering rebounds in viral counts. The new results also help explain the wide variability in patients' responses to a given treatment regime.

c) ENVIRONMENTAL MODELING: The program on Large-scale Computer Models for Environmental Systems had two main emphases, one on atmosphere/ocean models and the other on flows in porous media.

Within the atmosphere/ocean part of the program, one of the principal themes was spatial-temporal statistics. A focus of the effort was in creation of "physical-statistical models," which combine deterministic PDEs with stochastic elements to produce a hierarchical model, solved by Monte Carlo Bayesian techniques. New computational methods were developed for such models, with applications to air pollution and earthquakes. Another theme was turbulence: intermittency leads to seemingly stochastic phenomena, including heavy-tailed distributions and long-range correlations. Applied mathematicians interacted with statisticians to develop new diagnostic techniques in this area. Research also led to new understanding of sub-grid-scale processes, such as the effective diffusivity of a passive scalar in the presence of a time-varying shear profile.

Another area of research was source apportionment modeling, which solves inverse problems in atmospheric chemistry to find sources of a pollutant given measurements of its distribution. Research at SAMSI helped to identify the role of spatial statistics in this problem, with applications to the global distribution of carbon monoxide and to atmospheric ammonium compounds in the U.S. Other themes explored during the program included the use of personal exposure models to study the human health effects of pollution; the design of monitoring networks; the combination of data from different monitoring networks; climatological extremes; and the use of cyclostationary stochastic processes.

The porous media part of the program had three main sub-themes: fundamental mathematical modeling, based on recently developed families of conservation laws, for fluid flow through multiphase porous media; numerical solution techniques; and design and optimization problems in porous media systems. Theoretical developments led to new approaches to pore-scale models to develop closure relations, lattice-Boltzman modeling of closure relations, and a new theoretical framework for averaging theory based model development and closure. Developments in numerical methodology included a new spatially and temporally adaptive solution to Richards' equation; higher-order discontinuous Galerkin methods; and the development of a problem-solving environment to facilitate the rapid development of environmental models. Work in optimization focused particularly on the development of a set of Community Problems to compare a variety of optimization methods.

A unifying theme through both parts of the program was multiscale modeling; both statisticians and applied mathematicians were interested in using information gathered at one spatial and/or temporal scale to predict phenomena at another. The long-term product of this program will be new understandings of how microscopic phenomena affect processes over large scales and ultimately the global environment.

d) DATA MINING AND MACHINE LEARNING: The DMML Program of 2003-2004 significantly impacts the following scientific and societal issues.

*Controlling The Cost of Pharmaceuticals*: The cost of developing a new drug already exceeds $1 billion, and shows no sign of even growing more slowly, let alone falling. Compounding this is the knowledge that the pharmaceutical world of the future is not one of "one size fits all" blockbuster drugs, but one of niche drugs that target rare diseases, specific patient groups, or both. What is not lacking is data: pharmaceutical companies screen millions of chemical compounds per year. But, they don't do it very efficiently, and their methods are not able, for example, to distinguish multiple modes of activity for compounds. The DMML program has produced statistical models and software tools that can determine that a compound is active via multiple mechanisms, which is a critical step in the direction of drugs tailored to both diseases and patients.

*More is More:* One of the fundamental tenets in science has always been the preference for simple explanations over complicated ones. There is even a name for this principle: Occam's razor. For problems involving the analysis of data, Occam's razor translates to "parsimony:" the simpler the description, or model, of the data, the better. But how the simple description is arrived at makes a difference. Traditionally, a simple description comes from a complete---but only minimally complete---description of the data. The DMML program has shown how and why it can be better instead to start from an overcomplete, redundant description. When this is done properly, the resulting simple description is simpler and more insightful than ones derived in the traditional manner.

*Real World Complexity:* If automobile dealers were able to sell cars one day more quickly, then collectively across the country they would save more than $1 million per day in interest. But what models, with what colors, and equipped with which options, sell faster where? Using data provided by a major US automobile manufacturer (also a NISS affiliate), the DMML program has shown that very complicated data (3.5 million vehicles, with more than 700 options in more than 200,000 combinations) can be mined to yield insight into the problem, if not an immediate path to solution. For example, brand matters more than geography, and color is a much better predictor of how long it takes to sell a vehicle than the configuration of options. (Possibly Henry Ford's "Any color you want, as long as it's black" maxim was correct.) The study has also shown in a dramatic way the effectiveness (some would say the necessity) of using simple tools to analyze large, complex data sets.

e) INTERNET TRAFFIC MODELING: This program has generated three way interdisciplinary collaboration between statisticians, probabilists and computer scientists / engineers. While there have been pairwise collaborations in the past, this new three way collaboration was truly exciting. Evidence of success was a number of jointly authored papers spanning a continuum between statistical visualization, probability theory, and new lessons about and tools for network analysis.

The workshops were successful in engaging a large number of people in SAMSI. Topics covered in the workshops included Network Tomography, Sensor Networks and Heavy traffic Congestion Control. The first two workshop were jointly funded by SAMSI and the National Security Agency.

The program research effort got off to a fast start, using some creative ideas for workshop format. To keep all participants fully engaged, each workshop was only a day

and a half in length. Most were centered around theme problems, and the traditional talk-discussion format was replaced with other activities. Each workshop began with a "5 Minute Madness" Session, where selected participants gave presentations limited to 5 minutes. Speakers at this included most senior speakers, because everyone wanted to hear what they were doing, but also a cross-section of junior people which gave the audience a clear understanding of the breadth of interests present. The next event was typically Theme Problem presentations, by carefully chosen speakers and discussants, aimed at promoting thoughts and ideas about the theme problems. This was followed by breakout discussion groups. Participants were assigned to these groups, to ensure a healthy mix of both disciplines and ages in each group. A group size of 8-10 worked very well, with effective and interesting conversation among the more senior people, but also with an opportunity for junior people to listen, to contribute, and to ask questions if they were not following the discussion. Finally, the results of discussion group were reported back to full meeting, by a senior person previously selected as group leader. An interesting, perhaps surprising, phenomenon was that quite senior people (at the level of frequently giving plenary talks at other meetings) from all fields were interested to join these workshops without giving a plenary talk, and instead just speaking in the 5 Minute Madness Session and reporting on the results of their breakout discussion group.

f) MULTISCALE MODELING: While the program is still ongoing, some exciting advances have been made in two areas. The working group on *Paradigms for Bridging Scales* has focused on fundamental energy mechanisms for piezoceramic and magnetic materials employed in present and projected high performance applications. A magnetic prototype has been chosen, and the group is developing both deterministic and stochastic models which incorporate multiple scales ranging from microscopic to macroscopic. Current research is focused on the development of hierarchical models employing Markov-Chain Monte-Carlo constructs and comparing results to numerical data generated using state-of-the-art micromagnetic codes. The goal is to develop data-driven modeling frameworks for advanced materials which characterize hysteresis and nonlinear dynamics in a manner which facilitates control design. This is a collaboration between applied mathematicians, statisticians and material scientists, and the combined modeling approach is a first in this area of science.

The working group on *Control Design* had considered a number of nonlinear control issues pertaining to systems exhibiting the nonlinear, hysteretic behavior characteristic of present smart materials. A nonlinear hybrid method comprised of feedforward and feedback loops has been numerically implemented, and the real-time implementation of this design is being investigated. Another exciting advance is the development of stochastic models and control designs which incorporate concepts related to wavelet analysis and stochastic Kalman filtering, with the goal of providing robust control designs suitable for real-time implementation.

## 2. Human Resources and Education/Outreach

SAMSI's impact on human resources is fully discussed in sections I.B and I.C, with impact on diversity highlighted in section I.H. The individual program reports also contain significant insight into human resource development. The successes of the SAMSI courses

and education and outreach programs are fully discussed in section I.E.4 and appendix G. Here we simply give two illustrations indicating the unique impact that SAMSI has in these areas.

As an indication of the success with postdoctoral fellows in instilling the SAMSI philosophy of merging applied mathematics and statistics, consider the case of Yanyuan Ma. She came to SAMSI with a Ph.D. in applied mathematics with the determined intention to become equally expert in statistical methods. After the first year at SAMSI, she spent her second year as a CRSC postdoc (note that our standard is to secure two years of support for postdocs – one with SAMSI and another with a related research organization), working with statisticians (resulting in 5 publications) to improve her statistical contributions, while contributing to the HIV project (see appendix A). She has been very successful in both her contributions and her career objectives; indeed, she has just accepted a position, beginning Fall, 2004, as Assistant Professor of Statistics at Texas A&M University.

The new initiative this year in outreach to undergraduates was the development of one-and-a-half day workshops that move from the introductory to current research frontiers. For these workshops, the subject of data mining and machine learning (DMML) was chosen because these are among the most exciting recent developments in data analysis. But the issues, which span statistics, computer science, applied mathematics and domain science, are complex. Neither the issues nor strategies to address them are readily accessible to undergraduates who may know little more than calculus, elementary probability and Java programming. And although there is no OHSA to protect data miners in training, conveying the ideas without trivializing them is a challenge. The workshop developed to address this has been given twice to rave reviews, and a third version is planned for faculty at teaching institutions. The high-level "take" divides DMML into association rules (conditional probabilities), clustering (unsupervised learning) and classification (supervised learning for prediction). The DMML world is bigger and more diverse, of course, but this view is broad without being overwhelming. Applications contexts used to illustrate problems and solutions include software engineering, drug discovery and automobile sales data. Two recurring themes are "scalability, scalability, scalability" and KISS (Keep it simple, stupid.).

# E. Evaluation by the SAMSI Governing Board

(John Harer, Douglas Kelly, Jon Kettenring, Daniel Solomon – chair)

The Governing Board provides broad oversight for the Institute's administration, finances, and evaluation, and for relationships among the partnering institutions. As part of the annual evaluation, the Governing Board has elected to address four broad questions. That evaluation follows:

**1) Is the synthesis of applied mathematics with statistics occurring?**

The synthesis of applied mathematics with statistics is a central tenet of the SAMSI mission. There are notable examples of this synthesis in specific SAMSI programs, but the extent varies substantially across the full portfolio. Such synthesis is perhaps most evident in the programs on Multiscale Model Development and Control Design and on Inverse

Problem Methodology in Complex Stochastic Models. Each of these programs was cochaired by a statistician and an applied mathematician. Some features of the synthesis are described in the current report on the first program (section E.3) and the follow-up report on the second (appendix A). In addition to the scientific outcomes of these programs, we see evidence in the impact upon early career researchers. One notable example, mentioned in the synopsis in this executive summary, is a postdoctoral fellow, educated as an applied mathematician and who has recently been offered and accepted a faculty position in a Statistics Department. We see graduate students who, following involvement in the Inverse Problems Program, are undertaking thesis research combining advanced tools from both statistics and applied mathematics. We also see examples of the integration being promulgated still earlier along the career development path. For example, in connection with the program in Data Mining and Machine Learning, SAMSI has developed a workshop for undergraduates that moves from the introductory to current research frontiers. The workshop was given twice, to audiences that spanned mathematics, statistics and computer science. A third version is planned for faculty at predominantly teaching institutions.

**2) Is SAMSI science of high impact?**

Section D of the executive summary cites examples of the impact of SAMSI science. One such example is found in the Inverse Problems Program, where ideas from applied mathematics and statistics were merged to develop new techniques to use noisy, incomplete data to estimate parameters in complex HIV models. When combined with control theory techniques, the resulting models suggest how improved therapies---suboptimal controlled structured treatment interruptions---may be effective in permitting patients to stop therapy for long periods of time without suffering rebounds in viral counts. The new results also help explain the wide variability in patients' responses to a given treatment regime.

The lists of refereed publications associated with SAMSI programs (section I.G.) provide another measure of evidence of impact on the mathematical and disciplinary sciences. The impact of the applied mathematics/statistics synthesis on solving problems not amenable to either discipline alone, will take a more sophisticated analysis. Economic and societal impacts are also beginning to surface. For example, if automobile dealers were able to sell cars one day more quickly, then collectively across the country they would save more than $1 million per day in interest. Using data provided by an affiliated major US automobile manufacturer, the Data Mining and Machine Learning program has shown that very complicated data (3.5 million vehicles, with more than 700 options in more than 200,000 combinations) can be mined to yield insight into the problem of what vehicles with what features sell quickly in which locations. SAMSI has generated specific insights and more generic approaches that appear to be changing the way that the manufacturer thinks about the problem.

**3) Is SAMSI recognized and respected by the national (vs local) statistics and applied mathematics communities?**

As the Director notes in his summary, most of SAMSI's early programs were part of the original grant proposal and so featured local leadership. Programs now being planned

(section II) show a preponderance of leadership by scientists from outside SAMSI's partner institutions. The various participant lists for concluded programs provide ample evidence of the national and international draw of SAMSI activities. Other evidence is in the offers of partnerships with other organizations including the American Institute of Mathematics, the National Center for Atmospheric Research, the (Canadian) National Program on Complex Data Structures, as well as NSF's own MSRI and IPAM. A challenge remains to attract more Departments of Mathematics to the NISS/SAMSI University Affiliates program.

**4) Is the Directorate working?**

The directorate model serves SAMSI very well. A significant test of the structure is a pending transition in the directorate. The Governing Board is monitoring the change, and we expect the transition to be smooth. Among the strengths of the model is that there are clear divisions of responsibility among the members of the directorate, and the incumbents have excellent working relationships. The Governing Board Chair and the SAMSI Director have a biweekly telephone conference at which administrative and personnel matters are regularly discussed and issues addressed where they have arisen. There is also excellent cooperation among the partner universities and NISS to ensure that obligations are met and that SAMSI continues to flourish.

# Table of Contents

# I. Annual Progress Report

The previous annual progress report was complete in all details only through April, 2003. Hence, we also report activities in Year 1 programs that occurred subsequently and were not itemized in the report. These Year 1 programs were *Inverse Problem Methodology in Complex Stochastic Models*, *Stochastic Computation*, and *Large Scale Computer Models for Environmental Systems;* there final reports are in Appendices A, B, and C, respectively.

# A. Program Personnel

## 1. Program and Activity Organizers

***Program Organizers***

| Program | Name | Affiliation | Field |
|---|---|---|---|
| Inverse Problem Methodology | Richard Albanese | AFRL, Brooks AFB | Medicine |
| in Complex Stochastic Models | H.T. Banks (Co-Chair) | NCSU | Applied Math |
| | Marie Davidian (Co-Chair) | NCSU | Statistics |
| *2002-2003 SAMSI Program* | Sarah Holte | Hutchinson | Biostatistics |
| | Joyce McLaughlin | Rensselaer Poly | Applied Math |
| | Alan Perelson | LANL | Biology |
| | George Papanicolaou | Stanford, NAC | Mathematics |
| | John Rice | UC Berkeley | Statistics |
| | Robert Wolpert | Duke | Statistics |
| Stochastic Computation | Merlise Clyde (Co-Chair) | Duke | Statistics |
| | Jean-Pierre Fouque | NCSU | Mathematics |
| *2002-2003 SAMSI Program* | Alan Gelfand | Connecticut | Statistics |
| | David Heckerman | Microsoft, NAC | CS and Stat |
| | Mark Huber | Duke | Probability |
| | Greg Lawler | Cornell | Probability |
| | Jun Liu | Harvard | Statistics |
| | John Monahan | NCSU | Statistics |
| | Michael Newton | Wisconsin Madison | Bioinformatics |
| | Scott Schmidler | Duke | Bioinformatics |
| | Mike West (Co-Chair) | Duke | Statistics |
| Large-Scale Computer Models | Mark Berliner | Ohio State | Statistics |
| for Environmental Systems | Montserrat Fuentes | NCSU | Statistics |
| | William Gray | Notre Dame | Geosciences |
| *Spring 2003 SAMSI Program* | Gabriele Hegerl | Duke | Meterology |
| | Sallie Keller-McNulty | LANL, NAC | Statistics |
| | C. Tim Kelley | NCSU | Applied Math |
| | Andrew Madja | Courant Institute | Applied Math |
| | Richard McLaughlin | UNC | Applied Math |
| | Cass T. Miller | UNC | Environment |
| | Doug Nychka | NCAR | Geostatistics |
| | Richard Smith (Co-Chair) | UNC | Statistics |

| | | | |
|---|---|---|---|
| Data Mining and Machine Learning | David Banks (Co-Chair) | Duke | Statistics |
| | Mary Ellen Bock | Purdue, NAC | Statistics |
| *2003-2004 SAMSI Program* | Jerome Friedman | Stanford | Statistics |
| | Alan F. Karr (Co-Chair) | NISS | Statistics |
| | David Madigan | Rutgers | CS and Stat |
| | William DuMouchel | AT&T | Statistics |
| | Warren Sarle | SAS Institute | CS and Stat |
| Network Modeling for the Internet | Kevin Jeffay | UNC | Computer Science |
| | James Landwehr | Avaya Labs | Statistics |
| *2003-2004 SAMSI Program* | John Lehoczky | Carnegie Mellon, NAC | Probability |
| | J. S. Marron (Co-Chair) | UNC | Statistics |
| | Ruth Williams (Co-Chair) | UC San Diego | Probability |
| | Walter Willinger | AT&T | Computer Science |
| | Donald Towsley | Massachusetts | Computer Science |
| Multiscale Model Development | M. Gregory Forest | UNC | Applied Math |
| and Control Design | Doina Cioranescu | U Pierre & Marie Curie | Applied Math |
| | Alan Gelfand (Co-Chair) | Duke | Statistics |
| *2004 SAMSI Program* | David Schaeffer | Duke | Mathematics |
| | Murti Salapaka | Iowa State | Applied Math |
| | Ralph Smith (Co-Chair) | NCSU | Applied Math |
| | Christopher Wikle | Missouri | Statistics |
| | Margaret Wright | NYU, NAC | CS and Applied Math |
| Education & Outreach Program | H.T. Banks (Chair) | NCSU | Applied Math |
| | Johnny Houston | Elizabeth City State | Math and CS |
| | Rachel Levy | NCSU | Mathematics |
| | J. Blair Lyttle | Enloe HS, Raleigh | Statistics |
| | Negash Medhin | NCSU | Mathematics |
| | Daniel Teague | NC Sch Math & Sci | Mathematics |
| | Wei Feng | UNC-Wilmington | Math and Stat |

## *Activity Organizers*

| **Activity** | **Name** |
|---|---|
| Inverse Closing Workshop -- *May 14-15, 2003* | H.T. Banks, Marie Davidian |
| Environment One-Day Workshop in Porous Media -- *May 16, 2003* | William Gray, Cass Miller |
| Environment Workshop on Spatio-Temporal Modeling -- *June 1-6, 2003* | Montserrat Fuentes, Doug Nychka, Richard Smith |
| SAMSI/CRSC Interdisciplinary Workshop for Undergraduates -- *June 9-13, 2003* | H.T. Banks, Negash Medhin |
| Stochastic Computation Closing Workshop -- *June 26-28, 2003* | Merlise Clyde, Mike West |

| | |
|---|---|
| SAMSI/CRSC Industrial Mathematical & Statistical Modeling Workshop for Graduates -- *July 21-29, 2003* | H.T. Banks, Pierre Gremaud, Alan Karr, Negash Medhin, Ralph Smith |
| Data Mining & Machine Learning Opening Workshop -- *September 6-10, 2003* | David Banks, Alan Karr |
| Internet Program Workshop on Internet Tomography & Sensor Networks *October 12-15, 2003* | J.S. Marron, Ruth Williams |
| Internet Program Workshop on Congestion Control & Heavy Traffic Modeling -- *October 31-November 1, 2003* | J.S. Marron, Ruth Williams |
| Undergraduate Two-Day Workshop on Data Mining: Handling the Flood of Data -- *November 14-15, 2003* | H.T. Banks, Alan Karr |
| Multiscale Model Development and Control Design Program Opening Workshop & Tutorials -- *January 17-20, 2004* | Alan Gelfand, Ralph Smith |
| Data Mining Workday on Support Vector Machines -- *January 28, 2004* | Marc Genton |
| Data Mining Workday on Theory & Methods & Large p, Small n Inference -- *February 4, 2004* | David Banks |
| Data Mining Workday on Bioinformatics -- *February 11, 2004* | Stan Young, Jackie Hughes-Oliver |
| Undergraduate Two-Day Workshop on Data Mining: Handling the Flood of Data -- *February 13-14, 2004* | H.T. Banks, Alan Karr |
| Multiscale Program Workshop on Challenges in Soft Matter Materials -- *February 15-17, 2004* | Greg Forest |
| Planning Workshop for the Random Matrices Program -- *February 29-March 1, 2004* | Jim Berger, Iain Johnstone |
| Hot Topics Workshop on Mathematical Sciences Research to Meet National Security Needs -- *April 1-2, 2004* | Alan Karr, Christopher Jones |
| Multiscale Program Workshop on Fluctuations and Continuum Equations for Granular Flow -- *April 16-17, 2004* | Robert Behringer, David Schaeffer |
| Data Mining and Machine Learning Closing Workshop -- *May 17-18, 2004* | David Banks, Alan Karr |
| SAMSI/CRSC Interdisciplinary Workshop for Undergraduates -- *May 31-June 4, 2004* | H.T. Banks, Negash Medhin |
| Network Modeling for the Internet Closing Workshop -- *June 24-26, 2004* | J.S. Marron, Ruth Williams |

## 2. Program Core Participants and Targeted Experts

For each of the major programs, the following tables present the key participants for the programs. The participants are categorized and coded as follows:

**D –** *Distinguished Lecturer* for the program.

**F** – *Faculty Release Person*, defined as an individual from a partner university of SAMSI who is accorded release time for participation in the SAMSI program; the cost-sharing value of this release time is indicated.

**G** – *Graduate Student*, receiving a research assistantship, in the indicated amount, from SAMSI.

**N –** *New Researcher*, receiving the indicated support (salary and fringe benefits) from SAMSI

**P** – *Postdoctoral Fellow*, receiving the indicated support (salary and fringe benefits) from SAMSI.

**PA –** *Postdoctoral Associate*, receiving the indicated support (salary and fringe benefits or reimbursement of expenses) from SAMSI

**T –** *Targeted Expert*, an individual with particular expertise that is felt to be needed for progress in key elements of program research. Such individuals are brought in for shorter intervals of time, for transference of expertise to the program participants.

**U** – *University Fellow*, a key program participant, visiting for a semester or year, whose primary support is via indicated cost-sharing from a partner university.

**V** – *Core Visitor*, an individual from outside the Triangle who plays a major role in the program activities, by either a lengthy visit to the program or repeated visits involving ongoing program research.

**Grey** - is used to indicate funds that are provided by partner university cost sharing.

*Note:* For visitors who have yet to visit SAMSI or who are still at SAMSI, dollar amounts in the tables below are the expense allotment for the visitor.

### *I. Data Mining and Machine Learning*

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Banks, David | M | Duke U | Statistics | F |

| | | | | |
|---|---|---|---|---|
| Bayarri, M.J. | F | U of Valencia | Statistics | V |
| Breiman, Leo | M | U of California-Berkeley | Statistics | T |
| Brooks, Atina | F | North Carolina State U | Mathematics | G |
| Chipman, Hugh | M | U of Waterloo | Statistics & Actuarial Science | T |
| Chu, Jen-hwa | M | Duke U | Statistics | G |
| Clarke, Bertrand | M | U of British Columbia | Statistics | U |
| Cutler, Adele | F | Utah State U | Mathematics & Statistics | T |
| DuMouchel, William | M | AT&T | | T |
| Fokoue, Ernest | M | SAMSI | | P |
| Friedman, Jerome | M | Stanford U | Statistics | DL |
| Garcia-Donato, Gonzalo | M | U of Castilla-La Mancha | Mathematics | PA |
| Genton, Marc | M | North Carolina State U | Statistics | F |
| Gleser, Leon | M | U of Pittsburg | Statistics | V |
| Goel, Prem | M | Ohio State U | Statistics | V |
| Hooker, Giles | M | Stanford U | Statistics | T |
| Hughes-Oliver, Jacqueline | F | North Carolina State U | Statistics | F |
| Khatree, Ravi | M | Oakland U | Mathematics & Statistics | V |
| Levina, Elizaveta | F | U of Michigan | Statistics | T |
| Liang, Feng | F | Duke U | Statistics | F |
| Lin, Xiaodong | M | SAMSI | | P |
| Liu, Fei | F | Duke U | Statistics | G |

| | | | | |
|---|---|---|---|---|
| Liu, Peng | M | North Carolina State U | Statistics | G |
| Liu, Regina | F | Rutgers U | Statistics | T |
| Madigan, David | M | Rutgers U | Statistics | T |
| Martin, Yvonne | F | Abbot Laboratories | | T |
| Mitchell, Tom | M | Carnegie Mellon U | Computer Science | T |
| Nobel, Andrew | M | U of North Carolina | Statistics | F |
| Palomo, Jesus | M | U Rey Juan Carlos | | PA |
| Paulo, Rui | M | SAMSI & NISS | | P |
| Remyala, Greg | M | Louisville U | Mathematics | V |
| Sun, Dongchu | M | U of Missouri | Statistics | V |
| Truong, Young | M | U of North Carolina | Biostatistics | F |
| Welch, William | M | U of British Columbia | Statistics | T |
| Zhang, Helen | F | North Carolina State U | Statistics | F |
| Zhang, Tong | M | IBM | | T |
| Zhu, Ji | M | U of Michigan | Statistics | T |

# II. Network Modeling for the Internet

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Buche, Robert | M | North Carolina State U | Mathematics | F |
| Cleveland, William | M | Bell Labs | Statistics Research | T |
| Dinwoodie, Ian | M | Tulane U & Duke U | Statistics | U |
| Ghosh, Arka | M | U of North Carolina | Statistics | G |
| Harrison, J. Michael | M | Stanford U | School of Business | T |
| Hernandez-Campos, Felix | M | U of North Carolina | Computer Science | G |
| Jeffay, Kevin | M | U of North Carolina | Computer Science | F |
| Kurtz, Thomas | M | Wisconsin U | Mathematics & Statistics | DL |
| Maulik, Krishanu | M | EURANDOM | | PA |
| Michailidis, George | M | U of Michigan | Statistics | V |
| Nowak, Rob | M | U of Wisconsin | Electrical & Computer Engineering | T |
| Park, Cheolwoo | M | SAMSI | | P |
| Park, Juhyun | F | U of North Carolina | Statistics | G |
| Pipiras, Vladas | M | U of North Carolina | Statistics | F |
| Resnick, Sidney | M | Cornell U | Operations Research & Industrial Engineering | T |
| Riedi, Rolf | M | Rice U | Statistics and Electrical & Computer Eng | T |
| Rolls, David | M | SAMSI | | P |
| Stoev, Stilian | M | Boston U | Mathematics | G |

| | | | | |
|---|---|---|---|---|
| Taqqu, Murad | M | Boston U | Mathematics | U |
| Towsley, Don | M | U of Massachusetts | Computer Science | T |
| Veitch, Darryl | M | Sprint Labs | | T |
| Williams, Ruth | F | U of California-San Diego | Mathematics | T |
| Willinger, Walter | M | AT&T Research | | T |

## *III. Multiscale Model Development and Control Design*

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Chapman, Jonathan | M | Oxford U | Mathematics | DL |
| Davis, Jimena | F | North Carolina State U | CRSC | G |
| Ellwein, Laura | F | North Carolina State U | CRSC | G |
| Ernstberger, Jon | M | North Carolina State U | CRSC | G |
| Gelfand, Alan | M | Duke U | Statistics | F |
| Gleser, Leon | M | U of Pittsburg | Statistics | V |
| Grove, Sarah | F | North Carolina State U | CRSC | G |
| Krener, Arthur | M | U of California-Davis | Mathematics | U |
| Lada, Emily | F | SAMSI | | P |
| Lucas, Joseph | M | Duke U | Statistics | G |
| Mancini, Simona Cordier | F | U Paris 6 | Laboratoire J. L. Lions | PA |
| Newell, Andrew | M | North Carolina State U | CRSC | N |
| del Rosario, Ricardo | M | Philippines U | Mathematics | V |

| | | | | |
|---|---|---|---|---|
| Smith, Ralph | M | North Carolina State U | CRSC | F |
| Tjelmeland, Haakon | M | Norwegian U of Science & Technology | Mathematical Sciences | V |
| Vance, Eric | M | Duke U | Statistics | G |
| Zabaras, Nicholas | M | Cornell U | Mechanical & Aerospace Engineering | V |

## 3. Summary of Activity Participants*

| | | Underrepresented Groups | | |
|---|---|---|---|---|
| **Activity** | **# Participants** | **# Female** | **# African-American** | **# Hispanic** |
| Inverse Closing Workshop -- *May 14-15, 2003* | 26 | 9 | 1 | 1 |
| Environment One-Day Workshop in Porous Media -- *May 16, 2003* | 49 | 11 | 1 | 3 |
| Environment One-Day Workshop on Spatio-Temporal Modeling -- *June 1-6, 2003* | 73 | 15 | 0 | 2 |
| SAMSI/CRSC Interdisciplinary Workshop for Undergraduates -- *June 9-13, 2003* | 15 | 10 | 3 | 0 |
| Stochastic Computation Closing Workshop -- *June 26-28, 2003* | 43 | 11 | 2 | 4 |
| SAMSI/CRSC Industrial Mathematical & Statistical Modeling Workshop for Graduates -- *July 21-29, 2003* | 54 | 10 | 1 | 0 |
| Data Mining & Machine Learning Opening Workshop -- *September 6-10, 2003* | 105 | 24 | 4 | 3 |
| Internet Program Workshop on Internet Tomography & Sensor Networks -- *October 12-15, 2003* | 93 | 15 | 0 | 1 |
| Internet Program Workshop on Congestion Control & Heavy Traffic Modeling -- *October 31-November 1, 2003* | 62 | 7 | 1 | 0 |
| Undergraduate Two-Day Workshop on Data Mining: Handling the Flood of Data -- *November 14-15, 2003* | 30 | 20 | 5 | 0 |

| | | | | |
|---|---|---|---|---|
| Multiscale Model Development and Control Design Program Opening Workshop & Tutorials -- *January 17-20, 2004* | 92 | 18 | 3 | 1 |
| Data Mining One-Day Workshop on Support Vector Machines -- *January 28, 2004* | No Formal Registration | | | |
| Data Mining One-Day Workshop on Theory and Methods & Large p, Small n Inference -- *February 4, 2004* | No Formal Registration | | | |
| Data Mining One-Day Workshop on Bioinformatics -- *February 11, 2004* | No Formal Registration | | | |
| Undergraduate Two-Day Workshop on Data Mining: Handling the Flood of Data -- *February 13-14, 2004* | 21 | 12 | 2 | 0 |
| Multiscale Program Workshop on Challenges in Soft Matter Materials -- *February 15-17, 2004* | 68 | 12 | 0 | 1 |
| Planning Workshop for the Random Matrices Program -- *February 29-March 1, 2004* | No Formal Registration | | | |
| HOT TOPICS Workshop on Mathematical Sciences Research to Meet National Security Needs -- *April 1-2, 2004* | 32 | 2 | 0 | 0 |
| Multiscale Program Workshop on Fluctuations and Continuum Equations for Granular Flow -- *April 16-17, 2004* | 33 | 3 | 0 | 0 |
| Data Mining and Machine Learning Closing Workshop | scheduled May 17-18, 2004: will be reported next year | | | |
| SAMSI/CRSC Interdisciplinary Workshop for Undergraduates | scheduled May 31-June 4, 2004: will be reported next year | | | |
| Network Modeling for the Internet Closing Workshop | scheduled June 24-26, 2004: will be reported next year | | | |

*Participant lists for workshops are given in Appendix A.

# B. Postdoctoral Fellows

This section starts with a brief synopsis of the activities of each postdoctoral fellow and associate, with further details in the following sections. An overview of SAMSI activities and strategies for effective mentoring of Postdocs is given in Section B.1. The mentoring assignments for the postdoctoral fellows are summarized in Section B.2. The midyear activity reports, written by the postdocs that were at by SAMSI in the Fall of 2003, appear in Section B.3. Annual activity reports from the postdocs appear in Section B.4. The results of the SAMSI postdoc questionnaire, which is aimed at assessing the quality of the SAMSI Postdoc experience, and at directly soliciting information for improvement, appear in Section B.5. Final reports, which track the post-SAMSI experiences of the 2002-2003 postdocs are in Section B.6.

The SAMSI Postdoctoral Fellows, for 2003-2004, with a brief synopsis (details available in Sections B.3, B.4 and B.5 below) of their activities were:

**Ernest Fokoue**, (at SAMSI for the full 2003-2004 year) participated in the Data Mining and Machine Learning Program working groups on Support Vector Machines and Large p Small n, led the subgroup on Bayesian SVM. Author or co-author of two papers currently in preparation.

**Emily Lada**, (at SAMSI for the Spring of 2004) participated in the Multiscale Model Development and Control Design Program working groups on Control Design and Paradigms for Bridging Scales. Developed simulation model of nafion, and automated steady state simulation output analysis. Co-author of four papers currently under review.

**Xiaodong Lin**, (at SAMSI for the full 2003-2004 year) participated in the Data Mining and Machine Learning Program, working on Bioinformatics, Bayesians Effective Sample Size, Privacy Preserving Statistical Analysis, Feature Selection for SVM, and Dimension Reduction and Clustering. Co-author of eight papers at various stages of development.

**Cheolwoo Park**, (at SAMSI for the full 2003-2004 year) participated in eight of the nine working groups in the Network Modeling for the Internet Program, and the Support Vector Machine working group in the Data Mining and Machine Learning Program. Co-author of eleven papers at various stages of development.

**David Rolls**, (at SAMSI for the Fall of 2003) participated in the Network Modeling for the Internet Program working groups on Suites of Models, Multifractional Brownian and Stable Motion, Semi-Experiments Formal Testing, Testbeds – Lab Experiments. Co-author of eleven papers at various stages of development.

The SAMSI Postdoctoral Associates, for 2003-2004, with a brief synopsis (details available in Sections B.3, B.4 and B.5 below) of their activities were:

**Gonzalo Garcia Donato**, (associated with SAMSI for the Spring of 2004) worked on the NISS project on "Validation of Complex Computer Models", which will feed into the upcoming SAMSI Computer Modeling Validation Workshop. Co-author of eight papers at various stages of development.

**Murali Haran**, (associated with SAMSI for the full 2003-2004 year) participated in the Data Mining and Machine Learning Program working groups on Theory & Methods and Large p, Small n. Did research on software instrumentation and Bayesian approaches to data mining. Co-author of two papers in preparation.

**Jesus Palomo**, (associated with SAMSI for the Spring of 2004) participated in the Data Mining and Machine Learning Program working groups on Theory & Methods and Large p, Small n. Worked on the NISS project on "Validation of Complex Computer Models". Co-author of six papers at various stages of development.

**Rui Paulo,** (associated with SAMSI for the Fall of 2003) worked on the NISS project on "Validation of Complex Computer Models", which will feed into the upcoming SAMSI Computer Modeling Validation Workshop. Co-author of four papers at various stages of development.

## 2. Overview of Postdoc Activities and Mentoring Strategies

The SAMSI Postdoctoral Fellowship experience has continued to include opportunities for collaboration in the SAMSI spirit of bringing together Statisticians and Applied Mathematicians. These opportunities came during the SAMSI Workshops, during the Working Groups that met weekly at SAMSI, from the SAMSI courses, and from informal discussions and contacts.

The enhancement of contacts between SAMSI and NISS Postdoctoral Fellows, particularly those participating in different programs, as well as their contact with the SAMSI Directorate has continued to be a serious concern. Last year's monthly pizza lunch (together with a Postdoc research presentation) was viewed as successful, and has been broadened. This was done by separating the two major purposes into two separate events.

The actual lunch and informal discussion component was continued as a monthly pizza lunch. A new dimension was added in terms of planned discussion topics. These topics covered various aspects of academic folklore, i.e. "things every academic should know", of a type that too often doesn't arise in other conversations. A typical format was that the Directors took turns offering their (sometimes rather different) views and experiences, with frequent questions by the postdocs. Topics covered in this context included:

- The job search process, application (what should and should not be included), the selection process, interviews (good and poor strategies), the job offer system.
- The grant process, how to write proposals, how they are reviewed, a comparison of different scientific cultures.

- The publication process, writing papers, how the review process works, writing reviews, editorial decisions, choice of journals, cross-cultural differences.
- The academic promotion process, ranks, tenure, the review system.

The postdoc research presentation part of the former pizza lunch was turned into a bi-weekly Postdoc & Graduate Student Seminar Series. Graduate students were included as both audience members and speakers. A challenge that arose early on was that because of the rather diverse research projects underway, it was not always easy for people from very different research areas to stay interested. To address this issue, we adapted the format of "practice job interview talks". Because successful job interview talks are able to interest both experts and non-experts, this seemed like an ideal way to both come up with broadly accessible talks, and also to allow the speakers to practice this important skill. The experience was further enhanced by limiting each talk to 50 minutes, which left 10 minutes for discussion of both technical matters, and also presentation. We found that non-experts tended to give very helpful pointers about presentational points.

Effective mentoring of postdoctoral fellows continues to be an important SAMSI goal. A SAMSI mechanism for ensuring that each postdoctoral fellow had at least two people with whom they could personally discuss any concerns that might come up, was "double coverage" of mentoring assignments. This has been done by assigning both a "scientific mentor" (usually the senior scientist most connected with the research) and an "administrative mentor" (a member of the Directorate, different from the scientific mentor), to each postdoctoral fellow. The mentoring assignments for 2002-2003 are given in Section B.2. This mechanism was only partially effective last year, which we discovered from our Postdoc questionnaire, where it became clear that few of the postdocs knew who there administrative mentor actually was. This was addressed this year by playing up the role of the administrator at the beginning of the year, and by recommending meetings once a month, most of which were help in the conveniently remembered time slot immediately after the pizza lunch.

To assess performance of SAMSI in terms of the overall postdoctoral experience, a Postdoctoral Questionnaire was used in March 2003. This was an updated version of last year's questionnaire. The questions and answers from each postdoctoral fellow can be found in Section B.5 below. The single clearest impression from these is that overall the postdoctoral fellows were very happy with their SAMSI experience, and feel that it has given substantial value added to their careers. However, as expected, there was some variation in the experience.

As another means of assessing the quality of the SAMSI experience, the Scientific Mentors were asked to comment on each of the Postdoctoral Fellows. These reports are in Section B.6 below. Again the overall impression is very positive. It is clear that the Postdoctoral Fellows have made very important contributions to SAMSI. Another good indication is that the situation of the Postdoctoral Fellow with initial difficulties was recognized, and corrective steps were taken.

In summary, the SAMSI Postdoctoral Program has been generally very successful. The postdoctoral fellows have been making well appreciated contributions to their programs, and been gaining valuable career skills for themselves. As expected with any new program there are areas in need of improvement, but these have been identified, and plans for addressing them next year have been made.

## 2. Postdoctoral Fellow Mentoring Assignments

| | Scientific Mentor | Administrative Mentor |
|---|---|---|
| Ernest Fokoue | Alan Karr | James Berger |
| Emily Lada | Ralph Smith | H. T. Banks |
| Xiaodong Lin | Alan Karr | J. S. Marron |
| Cheolwoo Park | J. S. Marron | James Berger |
| David Rolls | Murad Taqqu | J. S. Marron |

## 3. Postdoctoral Fellows and Associates Mid-Year Activity Reports

These reports were written by each postdoctoral fellow or associate, in December 2003. Reports do not appear for postdocs who joined SAMSI in 2004.

**Ernest Fokoue**,
I am currently fully involved in two of the four Data Mining and Machine Learning working groups, namely the Support Vector Machine group and the Large p Small n group. As part of my activities within the SVM working group, I am the leader of the Bayesian SVM team, and as such, I have so far done the following:

(a) Collection of existing papers, references and software on the Bayesian treatment of Support Vector Machines,
(b) Test of existing methods on toy and real data
(c) Brainstorming with Professor Prem Goel on the statistical justification of the Relevance Vector Machine,
(d) Study of the possibility of unification of the various Bayesian approaches to SVM,
(e) Study of the relationship between design points in the Optimum Design literature and relevant points as presented in Relevance Vector Machine.

Considering the fact that the present implementation of the Relevance Vector machine is mainly empirical Bayes, I am considering a full Bayesian approach to Relevance Vector Machine, with an emphasis on the choice of "good" priors and the need to address some important computational difficulties inherent in the need to invert very large matrices.

Within the Large p Small n working group, I am currently developing a paper with Professor Bertrand Clarke. In this paper, we are proposing a novel method of function approximation based on sequential Model List Selection. Our aim in this work is to consider a variety of basis function sets and then sequentially refine and improve the approximation of the "true" function underlying a sample of observations. We use Predictive optimality as our measure of goodness, and we intend to use the scheme as a way to estimate the most appropriate level of overcompleteness required in a given setting. We are using Bayesian Model Averaging as our function estimator at each iteration of the sequential procedure. Model uncertainty being of particularly great interest to us, one of the future direction of this work is to study the formulation of a

generalized Bias-Variance decomposition that would serve as way to better quantify model uncertainty.

Besides the above mentioned groups, I am also indirectly working with the Bioinformatics team, more speci¯cally on the analysis of the Mono Amine Oxide (MAO) data. One of the key questions around the use of SVM on the MAO data is whether the euclidian kernels commonly used in SVM are appropriate for the MAO data whose descriptors are all binary. I am currently studying the use of non euclidian kernels for the MAO data on both the SVM and the RVM.

My ultimate goal is to gain insights into some properties of this type of data so as to design a kernel that would be a suitable measure of similarity in such settings.

**Xiaodong Lin**,

*1 Feature selection for SVM*

We suggest a new regularization method for variable selection in SVM. The idea is to replace the lasso-type L1 penalty by a nonconcave penalty called SCAD (smoothly clipped absolute deviation), proposed by Fan and Li (2002). We compare the results with the Newton method, by Bradley and Mangasarian (1998), Fung and Mangasarian (2002), and the feature selection for SVM proposed by Weston et. al. (2001). This is joint work with Helen Zhang, Jeongyoun Ahn and Cheolwoo Park [6].

*2 Analysis on Metabolism data set*

We investigate the use of robust singular value decomposition, rSVD, and recursive partitioning on metabolomic data set. Clustering using the metabolite data is only modestly successful. Using distances generated from multiple tree recursive partitioning was more successful. This is joint work with Susan J. Simmons, Chris Beecher, Young Truong and S. Stanley Young[4].

*3 Privacy preserving statistical analysis*

Privacy and security considerations can prevent sharing of data, derailing data analysis. We study the problem of privacy preserving statistical analysis which prevents disclosure of individual data items or any results that can be traced to an individual site. We have finished the work on privacy preserving linear regression over horizontally partitioned data set. This is joint work with Alan Karr, Jerome Reiter and Ashish Sanil [1].

*4 Effective sample size*

The problem is, based on the reference densitu, to find the sample size which minimizes the expected Kulback-Liebler distance over the posterior distributions. Currently we are proving the existence and uniqueness of a solution. Meanwhile, three examples which have implications in gene expression n data analysis and other application fields are studied. This is ongoing work with Bertrand Clarke [5].

*5 Mixture of factor analyzers and Degenerated EM algorithm*

This is continuing work from my Ph.D. thesis. We proposed the constrained mixture of factor analyzers model for simultaneous dimension reduction and clustering. The model contains analysis for the EM algorithm in a constrained parameter space and dynamic model selection based on a two-step iterative procedure. We study the likelihood

unboundedness problem for the Gaussian mixture models. A Degenerated EM algorithm is proposed to solve the EM breakdown problem on the boundary of the parameter space. These are joint work with Yu Michael Zhu [2] [3].

***References***

[1] Alan F. Karr, Xiaodong Lin, Jerome P. Reiter and Ashish P. Sanil. Secure Regression on Distributed Databases. To be submitted.
[2] Xiaodong Lin and Yu Michael Zhu. Degenerated Expectation Maximization for Local Dimension Reduction . Submitted.
[3] Xiaodong Lin and Yu Michael Zhu. Constrained Mixture of Factor Analyzers with Model Se-lection. In preparation.
[4] Susan J. Simmons, Xiaodong Lin, Chris Beecher, Young Truong and S. Stanley Young. Active and Passive Learning to Explore a Complex Metabolism Data Set. Submitted.
[5] With Bertrand Clarke. Bayesian Effective Sample Size. In preparation.
[6] With Jeongyoun Ahn, Cheolwoo Park and Helen Zhang. Variable Section for SVM using Non-concave Penalty. In preparation.

**Cheolwoo Park**,

I am currently a member of two programs of Statistical and Applied Mathematical Sciences Institute (SAMSI). One is “Network Modeling for the Internet” and the other is “Data Mining and Machine Learning”.

1. *Network Modeling for the Internet* (http://www.samsi.info/200304/int/int-project.html)

We have several working groups and I am now deeply involved in these groups. Our main theme problem is to characterize burstiness of Internet traffic and find the causes in order to build models that can mimic real traffic. This study will aid improvements in network components (e.g. switches) and protocols (e.g. Transmission Control Protocol (TCP)). To achieve this goal, exploratory analysis tools and statistical tests are needed, along with new models for aggregated traffic.

The main data we have used are packet and byte counts measured at the UNC main link in 2002 and 2003**.** One way to look at burstiness is to estimate the Hurst parameter, H, of the traffic, which is related to long range dependence analysis. We have been estimating H with different methods and summarizing our results at

http://www-dirt.cs.unc.edu/net_lrd/.

Based on these results, we are preparing a paper, “Long range dependence analysis of Internet traffic” (with F. Hernandez-Campos, L. Le, J. S. Marron, J. Park, V. Pipiras, F. D. Smith, R. L. Smith, M. Trovero, and Z. Zhu).

SiZer (Significance of Zero crossings of the derivative) analysis, proposed by P. Chaudhuri and J. S. Marron, is a new visualization method to enable statistical inference to discover meaningful structure within data, while doing exploratory analysis using statistical smoothing methods. Using a dependent data version of SiZer, I demonstrated statistically significant differences between the data and fractional Gaussian noise (FGN) which has been a popular long range dependent model for Internet traffic. J. S. Marron, V. Rondonotti, and I are preparing a paper “Dependent SiZer: goodness of fit tests for time series models” based on the results at

http://www.unc.edu/%7Ecwpark/SiZER/SiZER_View.html.

Investigation of full wavelet spectra is one technique to look at scaling behavior of the traffic. For an FGN, a wavelet spectrum shows a straight line with a slope related to H. However, analyzing wavelet spectra of UNC traces, I found interesting features, including nonstationary behavior and changes in the traffic "fingerprint" between 2002 and 2003. Also, we are developing a tool for local analysis of self-similarity by adapting a wavelet method. This study was motivated by the fact that global self-similarity analysis of a long time series, for example H estimation of a full trace, reveals some limitations and ignores local nonstationary behavior. Regarding this topic, two papers are in the progress with M. Taqqu, S. Stoev, and J. S. Marron. One is "Strengths and limitations of the wavelet spectrum method in the analysis of Internet traffic" and the other is "Local analysis of self similarity in Internet traffic". Also, multi-scale analysis of Internet traffic is under develpoment with F. Hernandez-Campos, F. D. Smith, J. S. Marron, and D. Rolls.

Another statistical tool I am developing now is Wavelet SiZer, which combines wavelet coefficients and SiZer. A wavelet spectrum shows scaling behavior of a time series, but it lacks location information in the time domain. SiZer itself is a good tool to find locations of unusual behavior, but sometimes it fails to detect them in the case of a long time series. Wavelet SiZer is a powerful tool to detect hidden nonstationary behavior in the data and give exact location information. It displays wavelet coefficients at different scales and uses SiZer to determine where and how they are significantly different from what we expect. Some results of applying this method can be found at http://www-dirt.cs.unc.edu/net_lrd/wavesizer.html. This tool is combined with SiNos (Significant Non-stationarities) proposed by L. R. Olsen and F. Godtliebsen, and a paper "Wavelet coefficient based visualization and inference" is being prepared with F. Godtliebsen, M. Taqqu, S. Stoev, and J. S. Marron.

In real traffic, a set of packets passing between the same two end points that can be naturally grouped together, such as those of TCP connections, are said to form a flow. Flows are an important concept in the understanding of network traffic structure. To understand the structure of point process of flow arrivals, we have adopted the notion of semi-experiments proposed by N. Hohn, D. Veitch, and P. Abry. An example of a semi-experiments is the random reordering of blocks of a time series to modify the correlations of the data whilst preserving the original structure within blocks. Semi-experiments allow us to track down the connections and origins of scaling behavior. So far, a relationship between packets and flows has been found. Our current goal is to look at a relationship between flows and HTTP documents (aggregation of flows) through semi-experiments. I am analyzing several UNC flows with wavelet spectra and also doing some simulations to mimic traffic behavior (http://www-dirt.cs.unc.edu/semiexps/). D. Veitch, J. S. Marron and I are in the process of doing some analyses related to this topic.

Another interesting line of research is related to the joint population structure of sizes and rates of Internet flows. We proposed the Extremal Dependence Measure (EDM) to see this relationship

http://www.cs.unc.edu/Research/dirt/proj/marron/ExtremalDependence/

But our conclusion was contradictory to Zhang et al.'s. However, by considering a global family of data thresholds, I showed that the differing results are driven by different types of thresholding, and EDM is robust to these thresholdings. We are now

investigating the effect of thresholdings and testing EDM with various data sets (http://www-dirt.cs.unc.edu/NetDepend/). This work is under development with F. Hernandez-Campos, J. S. Marron, F. D. Smith, and K. Jeffay.

Internet micro burst phenomenon in terms of arrival times and sizes of Internet flows (with Z. Zhu and H. Shen), and a new algorithm using wavelet coefficients for estimating the Hurst parameter (with J. Park) are also my additional topics.

2. *Data Mining and Machine Learning*

I am participating in another SAMSI program, Data Mining and Machine Learning, especially in the SVM group (http://www.samsi.info/200304/dmml/web-internal/svm/svm.html). In pattern recognition, the problem of selecting relevant variables is important. For example, with microarray data, gene selection is a fundamental issue in gene expression-based tumor classification. Optimal subset selection is attractive because it provides simple and interpretable models, but it involves a combinatorial procedure and is known to be unstable. SVM has been extensively used as a classification tool with a great deal of success from object recognition. Recently, several feature selection algorithms for SVM have been proposed and we are also developing a new algorithm by applying a Smoothly Clipped Absolute Deviation (SCAD) penalty. The SCAD penalty was proposed by J. Fan and R. Li, and it attempts to automatically and simultaneously select variables. In addition, it achieves three desired properties: unbiasedness, sparsity, and continuity. We are expecting this new feature selection algorithm to provide better performance in both selecting variables and classification. This algorithm is under development with H. Zhang, J. Ahn, and X. Lin.

**Rui Paulo**,

The workshop that marked the end of the Stochastic Computation Project was held on June 26{28, 2003, and I presented a talk entitled \Bayes Factors and Marginals Likelihoods." This communication essentially summarized the findings of the Model Selection Group, lead by Professor Merlise Clyde, while studying the problem of comparing stochastic computation methods of characterizing the posterior distribution on the model space. Specifically, we observed that, at least in particular examples, importance sampling can present considerable advantages over modern, more sophisticated methods. We are in the process of producing a manuscript on this subject, in particular we are working on further examples that illustrate this point.

I spent the month of July as a visiting postdoc at the Universidad Politecnica de Cartagena, Spain, collaborating with Professor Mathieu Kessler on the problem of specifying objective priors for the parameters of Gaussian processes. This collaboration was made possible by the support of the European Research Training Networks.

At the Joint Statistical Meetings, held in San Francisco, CA, in August 3{7, 2003, I gave a talk entitled “Default priors for Gaussian processes having separable correlation structure” as a finalist of the Savage Award, Theory and Methods.

As a joint NISS/SAMSI postdoctoral fellow, I have spent more time involved in SAMSI projects during the 2002/03 academic year. Since August 2003, the situation is now reversed, and I have been devoting most of my time to NISS and the research project that deals with the statistical analysis and evaluation of complex computer models.

In particular, I have produced some software which implements a simplified Bayesian analysis of the framework the group has been developing. This is still research software, but it is expected to evolve over time into a serious application. We have decided to call this prototype version SAVE, which stands for Simulator Analysis and Validation Engine.

I have also given some minor contributions to the analysis of another testbed example, the Road Load Analysis project, which deals with a simulator of the behavior of shock absorbers in cars. Simultaneously, I have been devoting some time to finishing up and submitting papers. I have submitted a paper entitled "Default priors for Gaussian processes," which I have also listed as NISS technical report #139. I am in the process of revising a technical report I have listed as ISDS discussion paper 03{27, and I should be submitting it soon for publication. Its title is "Conditional frequentist sequential tests for the drift of Brownian motion." I am grateful to Professor James Berger for guidance through the manuscript preparation and submission process.

**David Rolls**,

*Papers Written*

1. Queues as a metric for Internet data, with George Michalidis, Felix Hernandez-Campos (in progress)

*Activities for Semi-experiment workgroup*

1. Performed trace-driven queuing analysis on over 25 traces http://www.samsi.info/200304/int/work/semiexps/semi_trace_queueing/index.html
2. Created webpage to make trace-driven queuing analysis accessible
3. Wrote computer programs to implement two kinds of "semi-experiments"
4. Performed "semi-experiments" on large (2+ GB) UNC datasets
5. Created webpages to make semi-experiment data accessible http://www.samsi.info/200304/int/work/semiexps/experiments/Semi-experiment_Data.html

*Activities for Testbed Workgroup*

(with George Michalidis, Don Smith, Felix Hernandez-Campos)

1. Performed queuing comparisons of original and replayed versions for two datasets, three different ways to account for mean trends
2. Created four webpages to present the results. http://www.samsi.info/200304/int/work/testbed/abilene1/queue_abiline1.html

*Presentations*

1. Some Basics of Queue Length, to SAMSI 'Semi-experiments' workgroup, Nov. 11, 2003
2. Animated Visualization of Burstiness, SAMSI Workshop on Congestion Control and Heavy Traffic Modeling (Oct. 31, 2003)
3. More Modeling and Better Models, to SAMSI 'Suite of Models' workgroup, Oct. 28, 2003
4. Three five-minute presentations throughout the fall semester

*Workshops Attended*

1. Workshop on Congestion Control and Heavy Traffic Modeling Oct. 31 – Nov. 1, 2003
2. Workshop on Internet Tomography, Oct. 12-14, 2003
3. Workshop on Sensor Networks, Oct 14-15, 2003

*Internet Program Courses Attended*

1. Statistical Analysis and Modeling of Internet Traffic Data
2. Long Range Dependence and Heavy Tails

*Other Activities*

1. Read 18 journal and conference papers, and several book chapters
2. Attended two talks by SAMSI distinguished lecturers: Margaret Wright (Oct. 7), Jerome Friedman (Nov. 4)

## 4. Postdoctoral Fellows and Associates Year End Activity Reports

These reports were written by each postdoctoral fellow or associate, in March - April 2004.

**Ernest Fokoue**

I am currently working on some of the chapters of the SAMSI Data Mining and Machine Learning monograph. More specifically, I am writing up a section on Sequential Model List Selection for Function Approximation and a section on Variational methods in statistical learning. I will also be contributing a chapter or a section of a chapter on Bayesian Analysis of Support Vector Machines and Kernel Methods. Finally, I will be contributing to the bioinformatics chapter on the analysis of the MonoAmine Oxidase (MAO) data.

I have worked with Prem Goel on finding a rigorous statistical justification of relevant points in the context of the Relevance Vector Machine. To date, the main achievement in this regard is that many interesting and promising ideas for such a characterization have arisen as a result of our brainstorming sessions, and many of them will be made more concrete during future sessions with the finality of generating a paper entitled ”On some statistical properties of the relevance vector machine and related methods”.

I am currently working with Prem Goel and Dongchu Sun on finalizing the write-up of a paper proposing a new hierarchical structure for the Relevance Vector Machine. The main result in this regard is that the extended prior structure will make it possible to obtain a unique solution, thereby improving on the RVM that does not provide a unique solution. To date, the 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. Once the computations are done, we intend to add the theoretical justifications and submit the paper for publication.

As part of my independent work, I proposed a new method for finding a sparse representation of an approximating function. The method is intuitively appealing. Unlike the RVM that is empirical Bayes, this new method is 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 method applies to both kernel expansion and basis expansion. The method has been tested on many toy problems used by the authors of SVM and RVM and the results are very encouraging. Future Work on this project will mainly concentrate on establishing the Relationship between the prevalence of vectors and the Design points. Part of future orientations will be to carry out a careful methodological justification of the convergence of the Markov Chains involved in the method.

The novel method of function approximation based on sequential Model List Selection has now reached the stage of computations. Early results are very promising, and the future orientation is to write up the theoretical justifications that back our results. The final aim is to submit the manuscript to the Journal of the American Statistical Association or any other journal that will emerge as the right home for our ideas.

I have also been participating in the discussions on the mining of text data from the census bureau, and I hope to contribute some useful ideas on the use of Support Vector Machines for such data.

I will be attending the 2004 Meeting of the International Federation of Classification Societies. The theme this year is Classification, Clustering, and New Data Problems and I intend to participate along with other SAMSI researchers.

*Publications and work in progress*

Fokoue, E and Clarke, B (2004). Sequential Model List Selection for Function approximation. Manuscript complete, to presented at ISBA Chile and submitted to the Journal of the American Statistical Association or any other journal.

Fokoue, E (2004). Sparsity through Prevalence Estimation. work completed and ready to be submitted to the Journal of Machine Learning Research.

Fokoue, E (2004). Mixtures of Factor Analyzers: Their Place in Machine Learning, to be presented at Interface 2004 in the highlights of the SAMSI Data Mining Year session, and to be published in the proceedings. Work in Progress

Fokoue, E, Sun, D and Goel, P (2004). A New Hierarchical Prior Structure for the Relevance Vector Machine. Work in progress, to be submitted for publication very soon.

Fokoue, E and Khattree, R (2004). Model Uncertainty, Model Selection and Model Averaging for some Kernel Based Methods with applications to Text Mining. Work at the early stages, to be specified further.

Goel, P and Fokoue, E (2004). On some statistical properties of the Relevance Vector Machine and related methods. Work in progress

House, L, Fokoue, E, Banks, D (2004). Robust Methods for Multidimensional Scaling, Clustering and Multiple Regression. Work in Progress

Fokoue, E (2004). Characterization of Overcompleteness in Function Approximation. Work in Progress to be presented at the Joint Statistical Meetings in Toronto. Work in Progress

**Emily Lada**,

*Working Groups*

I joined SAMSI in January 2004 and I have been participating in two working groups for the Multiscale Model Development and Control Design program: Control Design and Paradigms for Bridging Scales.

*Research Problems*

I am currently assisting in the development of a simulation model of nafion. The polymer chains making up nafion are composed of a hydrophobic back bone with hydrophilic side chains. In order to develop an energetics model for the hydrophilic regions, it is necessary to estimate the material properties of the hydrophobic region. The bulk material properties of the hydrophobic region depend on how stretched out the chains are. This effect is modeled by tracking the distribution of the end-to-end distance for a large number of chains. Collaborators on this project are Ralph Smith (NCSU), Jessica Matthews (NCSU), and Lisa Weiland (VA Tech).

I am also working on a project to develop an automatic method for steady-state simulation output analysis. This method will address two problems commonly associated with analyzing data from steady-state simulations: the correlation problem and the initialization bias problem. Collaborators on this project are James Wilson (NCSU), Natalie Steiger (University of Maine), and Jeff Joines (NCSU).

*Papers*

Lada, E. K. and J. R. Wilson. 2004. A Wavelet-Based Spectral Procedure for Steady-State Simulation Analysis. *European Journal of Operational Research,* in review.

Lada, E. K., Joines, J. A., Steiger, N. M., and J. R. Wilson. 2004. Performance of a Wavelet-Based Spectral Procedure for Steady-State Simulation Analysis. *INFORMS Journal on Computing,* in review.

Lada, E. K., Joines, J. A., Steiger, N. M., and J. R. Wilson. 2004. Performance of a Wavelet-Based Spectral Procedure for Steady-State Simulation Analysis. *Proceedings of the 2004 Winter Simulation Conference,* in review.

Steiger, N. M., Lada, E. K., Wilson, J. R., Joines, J. A., Alexopoulos, C., and D. Goldsman. 2004. ASAP3: A Batch Means Procedure for Steady-State Simulation Analysis. *ACM Transactions on Modeling and Computer Simulation*, in review.

**Xiaodong Lin**,
During the SAMSI program year on data mining and machine learning, I participated in all of the four working groups. SAMSI has given me great opportunities to collaborate with leading researchers in the corresponding fields. The following are a list of research projects I conducted at SAMSI.

*1 Bioinformatics*
Analysis of metabolite data set. We investigate the use of robust singular value decomposition (rSVD), recursive partitioning, support vector machine and random forest on the metabolomic data set. Based on these analysis, we can build classification rules to separate the patients to different disease groups. Meanwhile, a set of important metabolites can be identified. These are joint works with David Banks, Chris Beecher, Adele Cutler, Leanna House, Susan Simmons, Young Truong and Stanley Young [2][8].
Leave-1-out cross validation on large dimension small sample size data. We study the validity of leave-1-out cross validation for large p small n data sets. Our objective is to identify and characterize the phenomenons where this cross validation method fails.
This is ongoing work with David Banks and Stanley S. Young.

*2 Bayesian effective sample size*
Consider a posterior density for a parameter given a fixed sample. Suppose we form a second posterior density for the sample parameter, based on a different model and data set. Then we can evaluate the relative entropy distance between these two posteriors. We minimize the relative entropy over the second sample. The result is the sample that makes the second posterior as close as possible to the first in an inferential sense. If the first model permits dependence in the data and the second requires independence, the optimization gives the effective sample size. We present several examples of this optimization to reveal the effect of the nuisance parameters, the hierarchical dependence and the effect of correlation among the sample. This is joint work with Bertrand Clarke [4].

*3 Privacy preserving statistical analysis*
Privacy and security considerations can prevent sharing of data, derailing data analysis. We study the problem of privacy preserving statistical analysis which prevents disclosure of individual data items or any results that can be traced to an individual site. In particularly, we study privacy preserving linear regression model under the context of horizontally and vertically partitioned data set. Currently our focus is on privacy issues for record linkage. These are joint works with Alan Karr, Jerome Reiter and Ashish Sanil [3][7].

*4 Feature selection for SVM*
We suggest a new regularization method for variable selection in SVM. The idea is to replace the lasso-type L1 penalty by a nonconcave penalty called SCAD (smoothly

clipped absolute deviation), proposed by Fan and Li (2002). It has been shown in the regression setting that the SCAD penalty achieves sparsity and unbiasedness of the estimates simultaneously, while the L1 type penalty functions results in biased estimates. By experimental studies using the gene expression data set and the metabolite data set, SCAD-SVM works very well in terms of classification error and selecting the important features. Two cross validation methods are proposed to select the tuning parameter. The first one is the V-fold CV, and the other one is the GACV proposed for standard SVM in Wahba, Lin and Zhang (2001). Current we are generalizing these ideas to nonlinear SVM similar to Fung and Mangasarian (2002). This is joint work with Helen Zhang, Jeongyoun Ahn and Cheolwoo Park [1].

*5 Dimension reduction and clustering*

This is continuing work from my Ph.D. thesis. We proposed the constrained mixture of factor analyzers model for simultaneous dimension reduction and clustering. The model contains analysis for the EM algorithm in a constrained parameter space and dynamic model selection based on an iterative procedure. We study the likelihood unboundedness problem for the Gaussian mixture models. A Degenerated EM algorithm is proposed to solve the EM breakdown problem on the boundary of the parameter space. These are joint work with Yu Zhu [5][6].

*References*

[1] Ahn, J., Lin, X., Park, C. and Zhang, H. (2004), Variable Selection for SVM using Nonconcave Penalty. *In preparation*.

[2] Beecher, C., Cutler, A., House, L., Lin, X., Truong, Y. and Young, S. S. (2004), Learning a Metabolomic Dataset with Random Forests and Support Vector Machines. *submitted*.

[3] Karr, A., Lin, X., Sanil, A. P., and Reiter, J.P. (2004), Secure Regression on Distributed Databases. *Submitted*.

[4] Lin, X. and Clarke, B. (2004), Bayesian Effective Sample Size. *In preparation*.

[5] Lin, X. and Zhu, Y. (2004), Degenerated Expectation Maximization for Local Dimension Reduction, *Proceeding of 2004 Meeting of International Federation of Classification Societies, to appear*.

[6] Lin, X. and Zhu, Y. (2004), Constrained Mixture of Factor Analyzers for Simultaneous Dimensional Reduction and Clustering. *In preparation*.

[7] Sanil, A.P., Karr, A., Reiter, J.P., and Lin, X. (2004), Privacy Preserving Regression Modelling via Distributed Computation. *Submitted*.

[8] Simmons, S. J., Lin, X., Beecher, C., Truong, Y. and Young, S. S. (2004), Active and Passive Learning to Explore a Complex Metabolism Data Set, *Proceeding of 2004 Meeting of International Federation of Classification Societies. to appear*.

**Cheolwoo Park**,

Program: Network Modeling for the Internet

Website: http://www.samsi.info/200304/int/int-project.html

Program: Data Mining and Machine Learning

Website: http://www.samsi.info/200304/dmml/dmml-home.html

Working groups that I have been engaged over the year
- Changepoints and Extremes
  : http://www.samsi.info/200304/int/work/chptextr.html
- Formulation of Suite of Model
  : http://www.samsi.info/200304/int/work/suite.html
- Multifractional Brownian and Stable Motion
  : http://www.samsi.info/200304/int/work/brownian.html
- Structural Breaks
  : http://www.samsi.info/200304/int/work/breaks.html
- Comparison of Hurst Parameter Estimators
  : http://www.samsi.info/200304/int/work/hurst.html
- Semi-Experiments
  : http://www.samsi.info/200304/int/work/semiexps/semi_look.html
- Testbeds – Lab Experiments
  : http://www.samsi.info/200304/int/work/testbeds.html
- SiZer and Wavelets
  : http://www.samsi.info/200304/int/work/sizerw.html
- Heavy Traffic
  : http://www.samsi.info/200304/int/work/heavy.html
- Feature selection in Support Vector Machine
  : http://www.samsi.info/200304/dmml/web-internal/svm/svm.html

4. Research Papers

1) “Visualization and Inference Based on Wavelet Coefficients, SiZer and SiNos”
- Status: submitted to Statistical Science
- Authors: Cheolwoo Park, Fred Godtliebsen, Murad S. Taqqu, Stilian Stoev, and J. S. Marron
- Technical Report Series: 2004-10
- Website: http://www-dirt.cs.unc.edu/net_lrd/wavesizer.html

2) “Long-Range Dependence in a Changing Internet Traffic Mix”
- Status: submitted to Computer Networks
- Authors: Cheolwoo Park, Hernandez-Campos, J. S. Marron, David Rolls, and F. D. Smith
- Technical Report Series: 2004-9

3) “Strengths and Limitations of the Wavelet Spectrum Method in the Analysis of Internet Traffic”
- Status: submitted to Computer Networks
- Authors: Stilian Stoev, Murad S. Taqqu, Cheolwoo Park, and J. S. Marron
- Technical Report Series: 2004-8

4) “LASS: a Tool for the Local Analysis of Self-Similarity”
- Status: submitted to Computational Statistics and Data Analysis

- Authors: Stilian Stoev, Murad S. Taqqu, Cheolwoo Park, George Michailidis, and J. S. Marron
- Technical Report Series: 2004-7

5) “Dependent SiZer: Goodness of Fit Tests for Time Series Models”
- Status: in progress (writing), invited to Journal of Applied Statistics
- Authors: Cheolwoo Park, J. S. Marron and Vitaliana Rondonotti
- Technical Report Series: 2004-11
- Website: http://www.unc.edu/~cwpark/SiZER/SiZER_View.html

6) “Long Range Dependence Analysis of Internet Traffic”
- Status: in progress (writing)
- Authors: Hernandez-Campos, Long Le, J. S. Marron, Cheolwoo Park, Juhyun Park, Vladas Pipiras, F. D. Smith, R. L. Smith, Michele Trvero, and Zhengyuan Zhu
- Website: http://www-dirt.cs.unc.edu/net_lrd/

7) "Semi-experiment analysis of the shifting knee wavelet spectrum"
- Status: in progress (analyzing)
- Participants: Cheolwoo Park, Darryl Veitch, Haipeng Shen, Felix Hernandez Campos, and J. S. Marron

8) “Thresholded Log-Log Correlation Analyses of TCP Response Characteristics”
- Status: in progress (analyzing)
- Participants: Cheolwoo Park, Felix Hernandez Campos, J. S. Marron, F. D. Smith, and Kevin Jeffay
- Website: http://www-dirt.cs.unc.edu/NetDepend/

9) “Robust H estimation, automatic choice of parameters”
- Status: in progress (analyzing)
- Participants: Cheolwoo Park and Juhyun Park

10) “Shot noise model, start times, micro-bursts”
- Status: in progress (analyzing)
- Participants: J. S. Marron, Cheolwoo Park, Haipeng Shen, and Zhengyuan Zhu

11) “Variable Selection for SVM using Nonconcave Penalty”
- Status: in progress (analyzing)
- Participants: Jeongyoun Ahn, Xiaodong Lin, Cheolwoo Park, and Helen Zhang

**David Rolls**,

Primary SAMSI Research:

a) Trace-driven Queueing Analysis
Collaboration with: C. Park, F. Hernandez-Campos, J.S. Marron, F.D. Smith

*Background:* Trace-driven queueing analysis is the idea of using a time series of packet or byte counts from a real network (a 'trace') and feeding it into a simulated queue. *Important* features in the trace are revealed by their influence on the queue. This is an attractive approach because the behavior of real queues in network switches and routers affects network performance.

Usual statistical approaches generally cannot distinguish which, or how, features affect queueing. My understanding of this point, and ability to perform trace-driven queueing analysis, was an unexpected strength that I brought to SAMSI, and my collaborations at SAMSI reflect this. I doggedly argued that queueing was the proper way to look at network traffic. This opened new investigations at SAMSI.

On request I performed trace-driven queuing analysis on over 20 traces from the UNC-CH main network link. This analysis required writing new programs in C, graphing with S-plus, and then creating web pages linked to graphics. The web pages are currently at http://www.samsi.info/200304/int/work/semiexps/semi_trace_queueing/index.html.

The utility of trace-driven queueing analysis led to my contribution of "Section 6" to the paper:
'Long-Range Dependence in Changing Internet Traffic Mixes', C. Park, F. Hernandez-Campos, J.S. Marron, D.A. Rolls, and F.D. Smith. 39 pages (submitted in 2004 to *Computer Networks*)

Usually trace-driven queueing simulation is used to compare similar traces (e.g. a real 'target' trace with a simulated version.) With this paper I had to devise a new way to compare unlike traces and then do the comparison.

b) Testbed Traffic Validation
Collaboration with: F. Hernandez-Campos, G. Michalidis, F.D. Smith
*Background:* The Computer Science department at UNC-CH has a testbed for network research. (It's really about 50 computers networked in a dumbbell shape with 25 on each end. Each computer pretends it's really, say, 500 computers, enabling simulation of a simple network of 25,000 computers.) The SAMSI 'testbed' workgroup was formed around the possibilities created by this hardware. But, they had a problem of how to verify that the network traffic they generated on the testbed looked sufficiently like real Internet traffic.

I doggedly argued that trace-driven queueing simulation could provide insight. I ultimately showed that their network traffic looked very different. The analysis required writing new programs in C, graphing with S-plus, and then creating web pages with linked graphics to report the results. The web pages are currently at
http://www.samsi.info/200304/int/work/testbed/abilene1/queue_abiline1.html.

Portions of this analysis, and complementary results obtained by George Michalidis, are included in the paper:
'Queueing Analysis of Network Traffic: Theoretical Framework and Visualization Tools', D. A. Rolls, G. Michalidis, F. Hernandez-Campos. 29 pages (submitted in 2004 to *Computer Networks*)

I expect another publication when they improve their method and we show their method works well from a queueing perspective.

c) Semi-experiments

Collaboration with: F. Hernandez-Campos, G. Michalidis, M. Devetsikiotis

*Background*: a semi-experiment (as used here) refers to the idea of isolating and changing one level of a multi-level model. (e.g. change the duration of network connections, but don't change the number of connections or their start times.) Two workgroups at SAMSI were created to explore the interesting idea of semi-experiments on network data. The goal is to reveal the relative importance, and suitability of various assumptions, for each level (e.g. is it *close enough* to assume connections arrive according to a Poisson process?)

I wrote computer programs to perform two semi-experiments, obtained suitable network files from F. Hernandez-Campos, overcame technical problems handling files larger than 2 GB, performed a semi-experiment multiple times on real network data and then made all the data available in a web page, currently at http://www.samsi.info/200304/int/work/semiexps/experiments/Semi-Experiment_Data.html.

This data will now be compared in various ways by people as they wish. My focus is a queueing comparison. New theoretical results are currently being investigated by G. Michalidis and I to do this comparison. This is ongoing work that should ultimately appear in a publication.

Other Activities:

Presentations

- Some Basics of Queue Length, to SAMSI 'Semi-experiments' workgroup, Nov. 11, 2003
- Animated Visualization of Burstiness, SAMSI Workshop on Congestion Control and Heavy Traffic Modeling, Oct. 31, 2003
- More Modeling and Better Models, to SAMSI 'Suite of Models' workgroup, Oct. 28, 2003

Workshops Attended

- Workshop on Congestion Control and Heavy Traffic Modeling Oct. 31 – Nov. 1, 2003
- Workshop on Internet Tomography, Oct. 12-14, 2003
- Workshop on Sensor Networks, Oct 14-15, 2003

**Gonzalo Garcia Donato**,

Began at SAMSI: January 22, 2004

*Working Groups*:

During my visit to SAMSI, i have been involved in the project "Validation of Complex Computer Models". This project is lead by Jerome Sacks, and the other collaborators are James O. Berger, Susie Bayarri, Rui Paulo, Jesus Palomo and Fei Liu. Within this group, I have worked in some multivariate generalizations of the proposed models.

*Research Problems:* I have started with Dongchu Sun some research in the Bayesian Model Selection area. More precisely, we study the characteristics of different families of `objective' prior distributions for the variance of the random effects. The results are promising and we have started written a paper which will be eventually submitted for publication.

*Production:*

-Gento, P, Ortega, J. and Garcia-Donato, G. (2004) ``Alternativas estadisticas al calculo del Valor en Riesgo," Estadistica Espanyola, 46, 155, 119-148.

-Garcia-Donato, G. and Chen, M.-H. (2004) ``Calibrating Bayes factors under prior predictive distributions," Statistica Sinica, to appear.

-Bayarri, M.J. and Garcia-Donato, G. ``A Bayesian, Sequential Look at u-Control Charts," submitted (revision stage) to Technometrics.

-Vallejos, R. and Garcia-Donato, G. ``Bayesian Analysis of Contaminated Quarter Plane Moving Average Models," submitted to Journal of Statistical Computation and Simulation.

-Garcia-Donato, G. and Bayarri, M.J. ``Divergence Based priors for objective Bayesian Model Selection," in progress.

-Garcia-Donato, G. and Bayarri, M.J. ``Conventional Selection of Normal Linear Models," in progress.

-Garcia-Donato, G. and Bayarri, M.J. ``Some relevant aspects of the Bayes fector," in progress.

-Garcia-Donato, G. and Sun, D. ``Objective priors for Random effects variances," in progress.

**Murali Haran**,
As a NISS postdoc, and a SAMSI associate for the academic year 2003-2004, I have benefited from the SAMSI program on Data Mining. I have actively participated in two of the Data Mining working groups - "Theory and Methods" led by Prof.David Banks, and "Large p, Small n", led by Prof.Bertrand Clarke. In addition to the weekly meetings and discussion groups, I have also attended the Data Mining tutorials, the opening 'kick-off' workshop, and the mid-term workshops, along with the semester long SAMSI course on Data Mining in the fall. Since my dissertation research was in Bayesian modeling and computation, being exposed to the problems and challenges of Data Mining has provided me with new perspectives and a new set of approaches to statistical modeling and problem solving. I have also learnt about research areas that I hope to work on in my academic career in the future. In addition, the SAMSI program has helped me build contacts with senior researchers and other young researchers, and the possibility of future

collaborations with them. Some of the classification techniques I have learnt from the Data Mining program are also proving to be useful for my NISS project on identifying failures in large software systems.

My publishable work is related to my NISS projects:

(1) "Instrumenting Software to Predict Failure" (in progress) with Alan Karr and Ashish Sanil.

(2) "A Bayesian model for relating browsing behavior to site structure on the World Wide Web" (in progress) with Alan Karr and Ashish Sanil.

**Jesus Palomo**,
During my visit to SAMSI as a Post-doc I have been involved in the following working groups within the Data Mining and Machine Learning program:

- Theory & Methods lead by David Banks.
- High dimensional Data (large p small n) lead by Bertrand Clarke.

In general, regarding both projects, my participation has been focused on helping to develop new methods for solving these particular problems. More specifically, I have started working with Prof. Banks (Duke Univ.) to develop new data mining methods for open source software projects - something that, eventually, could be sent to the special issue of the Management Science journal.

I have also been attending the Data mining course held at SAMSI.

I have participated in the research project entitled "Complex Computer Model Validation" lead by Jerome Sacks. Other participants in the working group are James Berger, Susie Bayarri, Rui Paulo, Danny Walsh, Fei Liu and Gonzalo Garcia-Donato.

I am currently developing the software that will be included in the forthcoming book "Objective Bayesian Inference" by James O. Berger, Jose M. Bernardo and Dongchu Sun. I am also designing the interface for this software package.

I have given the following talks:

- "Multiple testing in data mining", February 4th, 2004, at the Data mining and machine learning theory Midterm workshop, joint work with J. Berger and M.J. Bayarri.
- "Project cost forecasting: A Bayesian approach", March 11th, 2004, in the post-doctoral seminar sessions.

I am currently working on the following open problems:

- Finding reference priors for different multivariate distributions, with Prof. Berger (SAMSI and Duke U.) and Prof. Sun (Missouri U.).
- Multivariate approach in complex computer model validation, with James O. Berger, Rui Paulo and Gonzalo Garcia-Donato.

*Papers during my visit:*

Palomo, J., Rios Insua, D. and Ruggeri, F. (2003) " Expert's opinion and dynamic models with applications to project costing", SAMSI Technical report.

Palomo, J., Rios Insua, D. and Ruggeri, F. (2004) " Modeling external risks in project management", to be submitted to Management sciences Journal.

Palomo, J., Rios Insua, D. and Ruggeri, F. (2004) "Dynamic models with expert input with applications to Project Cost Forecasting", submitted to JASA (Applications section).

Palomo, J., Rios Insua, D. and Ruggeri, F. (2004) "A framework for bidding in procurement auctions", ERCIM News, num. 57. forthcoming April 2004.

Palomo, J., Rios Insua, D. and Ruggeri, F. (2004) "Asymmetric models for fist-price sealed-bids", in progress.

Salewicz, K., Rios Insua, D. and Nakayama, M. (2004) "Building the Bridge Between Reservoir Management and Decision Analysis", in progress.

**Rui Paulo**,

As I mentioned in the previous report, I have been spending some time finishing up and submitting papers. One paper, entitled “Default priors for Gaussian processes”, has been submitted to the Annals of Statistics. I have received a letter from the Editor stating that the paper has been tentatively accepted for publication, subject to suitable revision. The referee's and Associate Editor's reports did not raise any major questions, and hence I am confident that the paper will be published. I have submitted another manuscript to Biometrika. It is entitled “Conditional frequentist sequential tests for the drift of Browninan motion.” I have not received any letter from the Editor yet.

The research that stemmed from the work of the Model Selection group of SAMSI Stochastic Computation project is in the process of being shaped into manuscripts for possible publication. I have been interacting with Merlise Clyde and Feng Liang in the production of a paper entitled “Gaussian Hyper-Geometric and other Mixtures of g-Priors for Bayesian Variable Selection.”

I will be traveling to Chile for the 2004 International ISBA meeting and presenting work that was initiated also by the Model Selection Group. It has to do with a comparison of methods to compute marginal likelihoods and our observation that often a straightforward application of importance sampling can present advantages over modern and more complex methods. I will start working on this problem again soon, having in mind also the production of a manuscript.

The other component of my work is my involvement with the Complex Computer Model Evaluation group at NISS. We are at this stage working with a road load model data set, and I have been participating in its analysis. I have also been working in the general framework analysis, in particular in the production of general purpose software. I will be traveling to Banff, Canada this summer to present research developed within this group.

# C. Graduate Student Participation

## I. Data Mining and Machine Learning

**Atina Brooks** (Statistics, North Carolina State) participated in the Bioinformatics working group.

**Jen-hwa Chu** (Statistics, Duke) ran a series of simulation experiments (for the Large p, Small n working group) and did the graphics and other exploratory data analyses on the GM time-to-turn data. He gave presentations on both topics.

**Leanna House** (Statistics, Duke) was not supported by SAMSI but she was associated with the Theory and Methods working group. She worked on metabolomics and proteomics data, gave a presentation, did research on robustifying data mining methods, and helped David Banks to edit a refereed proceedings.

**Balaji Krishnapuram** (Electrical Engineering, Duke) was not supported by SAMSI, but he is a graduate student who does research in data mining. He gave a presentation on the use of unlabeled sample and a new approximation to EM calculations.

**Fei Liu** (Statistics, Duke) did outstanding work on overcompleteness and developing a minimum description length method for handling machine-state data of the kind that GM had intended to provide but did not-ultimately, we got a similar data set from IBM. She also did significant work on the use of unlabelled sample in classification and a review of EM techniques.

**Peng Liu** (Statistics, North Carolina State) led analysis of the NCAR space-time database by the SVM working group.

**Katja Remlinger** (Statistics, North Carolina State) was not supported by SAMSI but was a year-long participant in the Bioinformatics working group.

**Eric Vance** (Statistics, Duke) helped to develop a generalization of the Lasso, ridge regression, and other special techniques from an optimal programming perspective-Bridge Regression. He gave a presentation on this.

## II. Network Modeling for the Internet

**Arka Ghosh**, (Statistics & Operations Research, UNC), supported by SAMSI for the full 2003-04 year. Participated in all program workshops. Active member of Multifractional Brownian and Stable Motion, Semiexperiments – Formal Testing, SiZer and Wavelets, and Heavy Traffic working groups. Presented a Postdoc – Grad Student Seminar.

**Felix Hernandez Campos**, (Computer Science, UNC), supported by SAMSI for the full 2003-2004 year. Participated in all program workshops. Active member of Comparison of Hurst Parameter Estimators, Semiexperiments – Look and See, Testbeds – Lab Experiments working groups. Performed critical data base work, providing data for entire program.

**Myung Hee Lee**, (Statistics & Operations Research, UNC), supported by SAMSI for the Spring of 2004). Performed research on periodicities of flow arrival times within documents, and of round trip times. Started preliminary micro-array work, leading into Computational Biology Program next year.

**Chuan Lin**, (Operations Research, NCSU), not supported by SAMSI. Participated in Workshop on Heavy Traffic and Congestion Control. Active Member of Heavy Traffic working group.

**Juhyun Park**, (Statistics & Operations Research, UNC), supported by SAMSI for the 2003-04year. Participated in all program workshops. Active member of Changepoints and Extremes, Suite of Models, Comparison of Hurst Parameter Estimators, SiZer and Wavelets, and Heavy Traffic working groups, lead the Semiexperiments – Look and See working groups. Presented a Postdoc – Grad Student Seminar. Co-author on 2 papers in progress.

## III. Multiscale Model Development and Control Design

**Laura Ellwein** (Mathematics, NCSU): Abnormal cerebral blood flow can be an indication of cerebral vascular disease. Laura's research is concerned with the simulation of cerebral blood flow during postural change from sitting to standing. A system of ordinary differential equations is used for a nine-compartment model representing the systemic arteries and veins in the upper extremities, lower extremities, brain, and heart. The system is solved with Matlab using steady-state parameters found in literature for initial values.

Optimal control is used to adjust these parameters such that the model can be made to fit measured blood flow and pressure data. Physiologically based control mechanisms are used to describe how arterial and cerebral blood pressure drop during the posture change.

Future analysis will include the addition of time delays to analyze the timing of the onset and the duration of the control. In addition, we plan to investigate the effect of replacing one or several compartments with a more detailed one-dimensional (with one spatial dimension) fluid dynamics model based on partial differential equations.

Whereas Laura's research focuses on a biological application, it encompasses a number of multiscale and control facets. There is also inherent uncertainty which requires analysis similar to that under investigation for materials. Laura will additionally participate in the SAMSI Program on Computational Biology of Infectious Diseases in the Fall of 2004.

**Jon Ernstberger** (Mathematics, NCSU): Jon's research focused on the development of distributed parameter models (partial differential equation or PDE) for smart material uniforms operating in nonlinear and hysteretic regimes. In the first step of the model development, nonlinear constitutive relations are constructed by combining energy principles at the mesoscopic level with stochastic homogenization techniques to construct macroscopic models which characterize the hysteresis inherent to the materials. PDE based on these constitutive relations are constructed through conservation of force and momentum. The final step of the initial component of the investigation focused on the approximation of these PDE through Galerkin techniques to obtain finite-dimensional, vector-valued systems appropriate for simulation. The next step of this analysis will focus on the extension of these models and approximation techniques to more complex geometries (e.g., plates and shells) and alternative materials including magnetic compounds and shape memory alloys.

This research program fits very naturally within the context of the multiscale program since it encompasses aspects of multiscale model development and approximation for advanced material systems. The models developed in this investigation will employed for model-based control design for applications ranging from flow control to remote lens cleaning.

**Joe Lucas** (Duke): Joe is working in the Multiscale Modeling Program at SAMSI this semester. As part of this work, he attended the opening Workshop. He is currently attending the Tuesday evening class for which he is Latex'ing Eric Vance's notes. These are being provided through Jon Ernstberger to Ralph Smith to provide a written document for the course. Also, he is participating in the ``Paradigms for Multiscale Modeling" working group, for which he is attending meetings and keeping up with the reading.

Joe is studying the uses of partition of unity in simplifying local regression smoothing. Given the proper choice of weight functions, it can be shown that it is equivalent to splines. He wants to show that the technique can be arbitrarily close to loess as well (given a different choice of weight functions).

Joe has contributed strongly to both the working group and class. While coming from a statistics background, he has obtained a firm grasp of the deterministic energy relations used to construct nonlinear constitutive relations at the mesoscopic level. He is presently investigating the constructing of stochastic materials models based on inference principles.

**Eric Vance** (Statistics, Duke): This semester at SAMSI, Eric has been working in the Multiscale Model Development and Control Design program. His involvement consists of being a part of the working group which is studying paradigms for bridging scales. He attends the weekly working group meetings and reads various articles concerning the items we are discussing. Related to this work is the SAMSI course on Multiscale Model Development in which he takes notes for Joe Lucas to convert to Latex, and keep up with the readings and the homework assignments. Also, as part of his involvement with SAMSI, he is continuing research related to his work last semester in the {\bf Data Mining and Machine Learning} program. Eric's specific research includes developing a model of multiple latent factors in order to explain the covariance within metabolites for

diseased and non-diseased individuals. This work spawned from the ``Large P small n'' working group because the number of metabolites is larger than the number of individuals or data points observed.

Eric is also contributing strongly to both the working group and class. While his research does not directly pertain to multiscale model development or control designs, it provides perspectives which may prove advantageous for constructing hierarchical models for advanced materials.

## IV. Education and Outreach

Two SAMSI Graduate Fellows, **Jimena Davis and Sarah Grove**, participated in several SAMSI opening workshops (DMML and Multiscale), but their main activities were in assisting with the E&O program. During the fall semester, they assisted in the SAMSI sponsored NC-REN TV course by developing and testing new material for the course that combined statistical and computational methods to analyze data from the experiments. They also developed and tested software for new assignments for the class prior to it being assigned to the class. During the spring semester they have been developing and testing new material and software for the upcoming SAMSI/MAA PREP Workshop for college teachers to be held May 26-29, 2004 at the University of Louisiana, Lafayette. They will travel to this workshop and assist the SAMSI AD for E&O in presenting the material.

# D. Consulted Individuals

The individuals consulted for the broad selection of topics within programs and workshops were the members of two groups:

- The **Program Organizers**, listed in Section I.A.1.
- Members of the **Advisory Committees**, listed in Section I.J.

The specific topics that Program Working Groups chose to pursue were, in general, selected by the Working Group participants themselves, according to their combined interests. In almost all cases, however, a Program Leader headed each working group, so that specific research topics remained consistent with overall program goals. In Section I.E, the various Program Working Groups, and their members, are discussed.

# E. Program Activities

## 1. Program on Network Modeling for the Internet

### 1.1 Introduction, Motivation and Initial Ideas

Because of the size and complexity of the internet, and the nature of the protocols, Internet traffic has proven to be very challenging to model effectively. Yet modeling is critical to improving Quality of Service and efficiency. The main research goal of this program was to address these issues by bringing together researchers from three communities:

- Applied probabilists studying heavy traffic queueing theory and fluid flow models;
- Mainstream Internet traffic measurers/modelers and hardware/software architects;
- Statisticians.

The timing was deemed right for simultaneous interaction among all three communities, because of current trends away from dealing with Quality of Service issues through over-provisioning of equipment. This trend suggested that heavy traffic models would be ideally situated to play a leading role in future modeling of Internet traffic, and in attaining deeper understanding of the complex drivers behind Quality of Service. This SAMSI program aimed at catalyzing this process through building strong bridges among the three communities.

An additional goal was to enhance contact between these research communities and potential industrial partners. The location of SAMSI is ideal for this purpose, because the Research Triangle is becoming a world center for the networking industry.

**Program Leadership:** Kevin Jeffay (North Carolina), James Landwehr (Avaya Laboratories), John P. Lehoczky (Carnegie Mellon), J. S. Marron (North Carolina/SAMSI, Co-Chair), Ruth Williams (California San Diego, Co-Chair), Walter Willinger (AT\&T Research), Donald Towsley (Massachusetts).

### 1.2 Program Goals

Here are the goals stated in the Program's Goals and Outcomes document: the main value of the SAMSI Program on Network Modeling for the Internet will come from bringing together statisticians, probabilists and network researchers (computer scientists and electrical engineers). While there have been previous pairwise contacts between these groups, benefits are expected from strengthening these connections, and the three way collaboration is novel, and expected to yield large benefits.

Specific research emphases, where this new three way interaction is expected to be especially fruitful, include:

- Measurement and Modelling
- Heavy Traffic – Congestion Control
- Internet Tomography
- Sensor Networks

## 1.3 Workshops

One workshop was originally planned for each of the 4 areas of emphasis listed above. The schedule for the first two was seriously disrupted by Hurricane Isabel. This was handled by turning the Measurement and Modeling Workshop into a smaller meeting of local participants (which Darryl Veitch of Sprint Labs attended as well), because we had critical mass in this area. The Heavy Traffic – Congestion Control Workshop was postponed (because of insufficient local involvement to start activities).

The format of all workshops was aimed at maximizing long term involvement. Plans varied somewhat, depending on the nature of the topic, and on the current state of the desired collaboration. These contacts were weakest in the case of Sensor Networks, so the format was mostly the traditional talk + discussion type, but with non-expert discussants employed as often as possible. For the other workshops, at least some pairwise collaborations had already been established, so activities were deliberately aimed at maximizing engagement of participants. Some of the devices, with discussion, are:

i. Five Minute Madness Introductory Session. A session of 5 minute talks, intended to serve as an introduction. While there is insufficient time for all participants to speak, a cross-section of junior and senior people, from all three backgrounds, gave all participants a clear view of the diversity of people present at the workshop. Some emphasis in selection of people is given towards senior researchers, with the idea that everybody would be interested in hearing a sound bite description of their current work (and there wouldn't be time for all of them to give longer talks).

ii. Theme Problems. These were intended to be the centerpoint of the workshop activities. They were be introduced by a very few carefully chosen talks, which were aimed at providing the basis of the following group discussion (as opposed to conventional talks on the speaker's current research). The theme problem for the cancelled Measurement and Modeling Workshop was to be "Causes of Burstiness". The theme problem for the Heavy Traffic Workshop was "Congestion Control, and what can be done in a network laboratory?". The theme problems for the Internet Tomography Workshop were "Validation of Tomographic Methods" and "Spatial Temporal data collection and analysis".

iii. Breakout Discussion Groups. These were the forum in which the main progress on theme problems was made. To maximize the desired three way collaborations, dsicussion groups were pre-assigned with the goal of

deliberate balance. Observation of previous SAMSI discussion groups revealed that when the groups were large, only a relatively few members actually contribute much to the discussion. To create an environment where young people felt comfortable joining the discussion and asking questions (thus feeling that they are also "owners" of the process), groups of size about 10 will be used. To ensure that discussion stays on topic, and to allow useful summarization of the results, each group had a leader (generally a senior scientist), who was responsible for reporting that group's results to the full group. To help leaders organize their thoughts, they were given a transparency and pen for use in the final reporting session.

iv. Pie in the Sky Session on big picture ideas. Here speakers were requested to give brief presentations on problems that they had no idea how to solve (in contrast to what they had recently done, or what they were planning to do next)

Intended lasting benefits, from these workshops, in terms of human resources included raising the awareness of participants of the value of our three-way collaboration, and encouraging their engagement in it. Researchers of all levels were targeted, but because it seemed likely the largest long term impact would be on young researchers (from all three areas, and from both academics and industry), a large share of resources was devoted to supporting their attendance. Because of the organizing committee's concept that this SAMSI program should also draw additional statisticians into the field of network analysis, a number of relatively young researchers (with an emphasis on under-represented groups) with no background in network analysis were invited. Our target group here was people who have been recently promoted, with the idea that they were most likely to be interested in (and perhaps even looking for) new research areas.

Detailed summaries of the workshops, including participants, schedule of events, and outcomes appears elsewhere.

### 1.4 Working Groups

The mini-meeting that replaced the Measurement and Modeling Workshop led to a discussion process aimed at developing ideas on how to organize this diverse group of people into effective working groups. A listing of topics of interest suggested a rough grouping as "Model Based Topics" and "Method Based Topics". A leader was selected for each group, who was charged with organizing and leading meetings, and eventually reporting to the larger group. To keep things straight, SAMSI Postdoc Cheolwoo Park constructed and maintained the program's web page at:
http://www.samsi.info/200304/int/int-project.html

This collaborative research had the largest direct impact in terms of human resources, because these were the people with the chance to most deeply feel the value of our three-way collaboration. As is clear from the group leaders listed below, from the above web page (which include all of senior and junior faculty, through postdocs, to

include even advanced graduate students), a wide variety of level of researchers was targeted and asked to join.

Details of these groups (Name, leader, members, group objectives) are:

1.4.1 *Changepoints and Extremes*, Richard Smith,

Robert Buche (North Carolina State University)
Fred Godtliebsen (University of Tromso, Norway)
Krishanu Maulik (EURANDOM, The Netherlands)
Cheolwoo Park (SAMSI)
Juhyun Park (University of North Carolina, Chapel Hill)
Haipeng Shen ( University of North Carolina, Chapel Hill)
Murad Taqqu (Boston University)
Haakon Tjelmeland (Norwegian University of Science and Technology, Norway)
Zhengyuan Zhu (University of North Carolina, Chapel Hill)

Pictures of bursty internet traffic tend to show a number of characteristics:
(a) periods of bursty behavior interspersed with non-bursty behavior;
(b) seemingly sharp transitions from one to the other (though it is not clear just how sharp these transitions are);
(c) within a bursty period, a much greater than usual frequency of extreme observations, both above and below the overall mean.
The objective of this working group is to explore approaches in which the transitions between bursty and non-bursty behavior are considered changepoints, with models and methods from extreme value theory used to characterize the bursty periods. The changepoints form some version of a point process (the simplest model would assume a homogeneous Poisson process, but others can be considered) while the bursty periods are viewed as independent realizations of some kind of threshold-exceedance process whose parameters are allowed to vary from one bursty period to another in the manner of a hierarchical model. The structure is flexible enough to allow for a variety of alternative specifications, and can be fitted through the Reversible Jump MCMC paradigm. Finally, it is proposed that the fitted model(s) be used to calculate a variety of "indexes of burstiness" (the development of which should be considered part of the research program) and thus used to compare different internet servers and paradigms.

1.4.2 *Formulation of Suite of Models*, Zhengyuan Zhu

Jay Aikat (University of North Carolina, Chapel Hill)
Kevin Jeffay (University of North Carolina, Chapel Hill)
Steve Marron (SAMSI/University of North Carolina, Chapel Hill)
Jonathan Mattingly (Duke University)
Krishanu Maulik (EURANDOM, The Netherlands)
Cheolwoo Park (SAMSI)
Juhyun Park (University of North Carolina, Chapel Hill)
Surajit Ray (University of North Carolina, Chapel Hill)

David Rolls (SAMSI)
Haipeng Shen (University of North Carolina, Chapel Hill)

Our goal is to develop useful statistical models for Internet traffic flow that are simple to analyze and simulate, and can capture the characteristics of actual traffic data that are important to electronics engineers and computer scientists. To achieve that goal, we will try to combine the top-down approach with the bottom-up approach. The class of statistical models we propose is based on the fact that Internet traffic is an aggregation of individual connections, each with a random starting time, random duration (from heavy-tail distribution) and a random throughput. We will use a bottom-up approach to model the starting time, duration, throughput and their dependence structure by analyzing the data derived from the actual traffic flow, and study the statistical property of the random process of aggregated packet counts thus obtained.
The working group will try to address the following problems:
1. Identify appropriate measures for evaluating how well a model fits the data (i.e., how to determin if model generated data is close to actual traffic data).
2. Find good models under those measures.
3. Develop methodology to estimate the parameters of the model as well as making statistical inference.
4. Simulate traffic flow efficiently under such model.
5. Other things group members want to address.

### 1.4.3 *Multifractional Brownian and Stable Motion*, Stilian Stoev

Robert Buche (North Carolina State University)
Arka Ghosh (University of North Carolina, Chapel Hill)
Krishanu Maulik (EURANDOM, The Netherlands)
George Michailidis (University of Michigan)
Cheolwoo Park (SAMSI)
David Rolls (SAMSI)
Robert Wolpert

The fractional Brownian motion (FBM) has been a very successful model for the traffic in modern telecommunication networks such as Ethernet-LAN and more generally, the Internet. FBM captures two major characteristic features of the network traffic: *time scale invariance* (statistical self-similarity) and *Long-Range Dependence* (LRD). A Gaussian stochastic process X={X(t)}, t>0 is said to be FBM if it has mean zero, stationary increments and is self-similar, that is, for all a>0, the processes {X(at)}, t>0 and a^H X(t), t>0, have equal finite-dimensional distributions. The parameter H belongs to the range (0,1) and is called the self-similarity parameter of the FBM process X. H is also the Hurst parameter of fractional Gaussian noise time series Y(k):=X(k)-X(k-1), k=1,2,... The FBM process can be also regarded as a physical traffic model. It appears in the limit of the superposition of independent ON/OFF sources with heavy-tailed ON and OFF periods, which mimic the flows in a busy network link. FBM is also the limit process of a physical infinite Poisson source

model with heavy-tailed sources. Current extensive studies of real traffic data, however, indicate that FBM alone cannot be used to explain the traffic burstiness (see http://www-dirt.cs.unc.edu/net_lrd/). Traffic burstiness appears to be a serious non-stationary effect in data, that cannot be contributed to seasonality or periodicity.

Our goal in this group will be to study, whether and how can the FBM model be augmented to account for non-stationarity effects (burstiness) in real data. We plan to focus on the so-called multifractional Brownian motion (MBM) model. The MBM processes $Y=\{Y(t)\}$, $t>0$ extend the class of FBM processes by allowing the self-similarity parameter H to change with time, that is, $H=H(t)$. More precisely, $Y(t)$ is defined by replacing the parameter H in an integral representation of the FBM procsess by a function of time $H(t)$, $t>0$. The resulting MBM process is Gaussian and locally self-similar, that is, $Y(t)$ behaves, locally, like a self-similar process with self-similarity parameter $H(t)$. We plan to investigate, whether these type of processes are relevant models for network traffic. More precisely, we will first focus on the following themes:

* Extend existing physical models (e.g. the infinite Poisson source model or the ON/OFF model) by obtaining the MBM as a stochastic-process limit.
* Explore the connections between the infinite Poisson source model and the ON/OFF model.
* Validation of the models under consideration, by:
  (a) developing techniques to estimate $H(t)$, locally, by using novel exploratory tools such as Dependent SiZer and wavelet analysis.
  (b) Estimating the parameters in the physical model from traffic data; experimenting with synthetic traffic data over a real network.

We look forward to actively collaborating with other workgroups on the above mentioned themes. The above listed topics are quite broad and preliminary, so any new related ideas and directions of research are very welcome. Please contact Stilian Stoev sstoev@samsi.info if you want to take part in this workgroup or any suggestions.

### 1.4.4 *Structural Breaks*, Vladas Pipiras

Robert Buche (North Carolina State University)
Fred Godtliebsen (University of Tromso, Norway)
Cheolwoo Park (SAMSI)
Stilian Stoev (Boston University)
Murad Taqqu (Boston University)

An increasing number of publications in the time series literature suggest that evidence of long-range dependence is an artifact of structural breaks. Structural breaks (also: level shifts, regime switching, structural instability) is one type of non-stationarity of a time series. The explanation for long-range dependence through structural breaks is particularly prevalent in Econometrics where evidence for long-range dependence was found in the time series of volatility of stock prices, inflation rates and other economic indicators. The goal of this working group is to explore the idea of structural breaks in the context of Internet traffic modeling where evidence of

long-range dependence is alsoubiquitous. It appears that these ideas have not been examined yet. The work of the group will focus around the following themes:
1. definition and types of structural breaks, models (some mixture, Markov-switching and other models),
2. structural breaks versus long-range dependence (possibly versus other phenomena),
3. statistical tests to discriminate between structural breaks and long-range dependence
4. applications to Internet traces.

### 1.4.5 *Semi-experiments - Look & See*, Juhyun Park

Ian Dinwoodie (Duke University)
Felix Hernandez Campos (University of North Carolina, Chapel Hill)
Kevin Jeffay (University of North Carolina, Chapel Hill)
Steve Marron (SMASI/University of North Carolina, Chapel Hill)
Jonathan Mattingly (Duke University)
Cheolwoo Park (SAMSI)
Surajit Ray ( University of North Carolina, Chapel Hill)
Haipeng Shen ( University of North Carolina, Chapel Hill)
Don Smith (University of North Carolina, Chapel Hill)

As a way of characterizing internet traffic, semi-experimental approach proposed by Darryl Veitch seems to lead us to another direction of thinking. In particular, it can provide an interactive tool that helps narrow a gap between modeling and data anlaysis. With a laboraroty at hand, this has a lot of potential to apply to our current study of traffic modeling. Main activities of our group will be to explore various aspects of traffic data by adapting semi-experimental approach and provides a ground for formal modeling.

### 1.4.6 *Semi-experiments - Formal Testing,* Mike Devetsikiotis

Ian Dinwoodie (Duke University)
Arka Ghosh (University of North Carolina, Chapel Hill)
Steve Marron (SAMSI/University of North Carolina, Chapel Hill)
Jonathan Mattingly (Duke University)
George Michailidis (University of Michigan)
David Rolls (SAMSI)
Zhengyuan Zhu (University of North Carolina, Chapel Hill)

The task of systematically characterizing the multitude of "factors" affecting traffic behavior and network performance, is a difficult but central one. Assessing the "effects" of such factors, ranging from qualitative statements, to full quantification, to (even better) sensitivity analysis, is particularly important. The approach of "semi-experiments" seems to present a refreshingly novel approach to this line of work. The objective of this work group is to engage in the careful and systematic study of such "semi-experiments" and attempt to

(a) formalize them in a statistical sense (i.e., relate them to design of experiments, factorial analysis, statistical confidence, hypothesis testing, etc.);
(b) combine them or extend them utilizing the group's own expertise in statistics, output analysis techniques, and network system design methodologies.

1.4.7 *Testbeds - Lab Experiments*, Don Smith,

Jay Aikat (University of North Carolina, Chapel Hill)
Felix Hernandez Campos (University of North Carolina, Chapel Hill)
Steve Marron (SAMSI/University of North Carolina, Chapel Hill)
George Michailidis (University of Michigan)
Cheolwoo Park (SAMSI)
David Rolls (SAMSI)
Haipeng Shen (University of North Carolina, Chapel Hill)

Networking research has long relied on simulation as the primary vehicle for demonstrating the effectiveness of proposed algorithms and mechanisms used in routers or TCP/IP protocols. Typically one constructs a network testbed in a laboratory and conducts experiments with actual network hardware and software (or one simulates network hardware and software in software such as the NS network simulator). In either case experimentation proceeds by simulating the use of the (real or simulated) network by a given population of users running applications such as FTP or web browsers. Traffic generators are used to inject application-level data objects into the network according to a model of how the applications or users behave. A critical aspect of this empirical methodology is ensuring that the resulting synthetic traffic, appearing as packets flowing through the network, preserves the essential characteristics of packet flows in real networks. An especially important property to study is the "burstiness" of packet-level traffic because it has been shown to have strong influences on many of the algorithms and mechanisms (e.g., active queue management) that networking researchers study. This working group could consider two important questions:
(1) How should we measure and characterize both real and synthetic packet-level traffic so we can verify that synthetic traffic preserves all the essential properties of real traffic?
(2) Can we design controlled experiments using a testbed network to confirm various hypotheses and findings from other empirical studies about the physical factors that lead to "burstiness" in traffic?

1.4.8 *SiZer and Wavelet*, Cheolwoo Park

Fred Godtliebsen (University of Tromso, Norway)
Arka Ghosh (University of North Carolina, Chapel Hill)
Juhyun Park (University of North Carolina, Chapel Hill)
Stilian Stoev (Boston University)
Murad Taqqu (Boston University)

In an analysis of long range dependent time series, a Logscale Diagram using a wavelet method is quite useful. Logscale Diagram is essentially a log-log plot of variance estimates of the wavelet details at each scale, against scale, complete with confidence intervals about these estimates at each scale. It can be thought of as a spectral estimator where large scale corresponds to low frequency. For example, one can estimate the Hurst Parameter from a Logscale Diagram by applying a weighted least square fit for a certain range of scales. SiZer enables meaningful statistical inference, while doing exploratory data analysis using statistical smoothing methods (e.g. histograms or scatterplot smoothers). It is a new visualization that brings clear and immediate insight into a central scientific issue in exploratory data analysis: Which features observed in a smooth of data are "really there"? This central question is critical in real data analysis, because discovery of a new feature, such as an unexpected "bump" or surprising "regions of decrease/increase", might lead to important new scientific insight. One common factor of these two tools is that they are looking the data at various scales. It is worth combining these two tools and make a new one for an analysis of long range dependent time series.

1.4.9 *Heavy Traffic*, Robert Buche

Arka Ghosh (University of North Carolina, Chapel Hill),
Chuan Lin (North Carolina State University)
Steve Marron (SAMSI/University of North Carolina, Chapel Hill)
Cheolwoo Park (SAMSI)
Juhyun Park (University of North Carolina, Chapel Hill)
Vladas Pipiras (University of North Carolina, Chapel Hill)

The Internet Heavy-Traffic working group has been working towards investigating open questions posed by Ruth Williams, UCSD, in her talk at a Modelling for the Internet Workshop and also described in the preprint by her and Frank Kelly "Fluid Model for a Network Operating under a Fair Bandwidth-Sharing Policy". In particular, the analysis for obtaining a workload reflected diffusion characterizing a part of a network operating at capacity ("heavy-trafffic") is needed. The working group is currently looking at relevant recent literature which will provide a motivation and structure for this analysis. The group meets every week or two and it is the hope that we will meet beyond the SAMSI program as it seems the interest of the group is strong.

## 1.5 SAMSI Courses

Two courses were taught during the Fall Semester of 2003. Both were listed at all 3 Triangle Universities, and were attended by a wide (both in terms of background, and also universities) range of students, both enrolled and auditing. Both courses met one evening per week, in the NISS main Classroom.

1.5.1: *Data Statistical Analysis and Modelling of Internet Traffic Data*
INSTRUCTOR: J. S. Marron

COURSE DESCRIPTION: The analysis and modelling of internet traffic data represents an important major challenge for engineers, for computer scientists, for statisticians and for probabilists. Really new ideas and models are needed because heavy tailed distributions and long range dependence (both appearing at a number of different points) render standard methods, such as classical queueing theory, unusable. This course considers a variety of mnethods for understanding and modelling internet traffic at a variety of levels, from individual TCP traces, to monitoring traffic on a main link. An important underlying concept is cross scale views of data. Novel graphical views of data play an important role. To reach a broad audience, prerequisites are kept to a minimum, with needed foundational material, including Q-Q plots, time series analysis, long range dependence, and SiZer analysis being introduced as needed.
COURSE WEB PAGE: http://www.samsi.info/200304/int/traffic-course.html

1.5.2: *Long range Dependence and Heavy Tails*
INSTRUCTOR: Murad S. Taqqu
COURSE DESCRIPTION: This course will focus on long-range dependence and heavy tails, notions which are relevant in computer traffic networks. Long-range dependence occurs when the covariances of a time series decrease slowly, like a power function. Heavy tails occur when the probability distribution of the time series has infinite variance and behaves like a power function. We will introduce self-similar processes which are idealized models that can encompass long-range dependence and/or heavy tails. We will focus first on fractional Brownian motion and on the related FARIMA time series models. To deal with infinite variance and heavy tails, we will introduce in a systematic fashion, infinite variance stable processes. We will study their properties and describe a number of stable (heavy-tailed) self-similar processes, including the so-called "Telecom model". We will also describe statistical methods for detecting the presence of long-range dependence and for estimating its intensity, focusing on wavelet methods since these are particularly useful in this regard.
COURSE WEB PAGE: http://www.samsi.info/200304/int/dependence-course.html

## 1.6 Continuing Collaboration

In addition to many new individual collaborations that have started (as indicated by co-authorship of the papers listed above), a larger scale effort is also currently under way, through a planned NSF IGERT proposal. Main partners on this proposal include many of the research partners during the SAMSI Program, and also a number of the most active participants during the Workshops.

The main goal of the IGERT proposal, named the Internet Statistics Education and Research Consortium (ISERC) is to continue to foster the special collaboration between statisticians, probabilists and network researchers that happened in the SAMSI program.

To achieve this goal, ISERC will depend upon leaders from all of these disciplines. Because the key players are geographically distributed, ISERC is envisioned as a Consortium (in contrast to the more common Center approach to collaborative research). The major activity of ISERC will be an annual research summer month, where travel expenses of all participants will be paid so they can come to NISS to work in an

atmosphere aimed at reproducing and multiplying the strong inter-disciplinary collaborations created by the SAMSI program. The SAMSI pilot effort has clearly proven the success of this approach in terms of both research and education.

As part of the educational component, ISERC cross-disciplinary course development will be encouraged (with progress reports featured at ISERC workshops). The challenge of doing this in a distributed fashion will be addressed using the Connexions collaborative approach to online learning (see http://cnx.rice.edu/), headed by ISERC Partner Richard Baraniuk.

Another goal of ISERC will be the collection, and web posting of high quality publicly available data sets. The value of such data sets, for a wide variety of purposes, became clear from several of the SAMSI workshops. Ongoing SAMSI activities include pilot data sets of this type, but there will be large value added from ISERC doing this on a much large scale (straightforward because of the geographical spread of the partners).

## 1.7 Program Assessment and Summary of Lessons Learned

In summary, the SAMSI Program on Network Modeling for the Internet was quite successful in terms of its main goal of establishing new interdisciplinary contacts. Less success was achieved on the goals of combining with industry, and internet posting of high quality data sets.

The collaborative efforts were a joy to behold. The excellent attitudes, and willingness to explore new research directions by the on site participants, gave a clear success in this direction. This was clearly the biggest success of the program.

The attempted contacts with industry took two major directions. Contacts with local industry fell through, perhaps because of the recent crash in the telecommuncations industry. There were massive layoffs (even complete company location closures) at a very important time, so people that were expected to be interested were too distracted by other matters. Contacts with nationwide industry were attempted through workshop attendance, and this generally much more successful. Industry attendees reported a very positive experience, but we were not successful in engaging them in on-going activites. It is not clear that alternative strategies could have helped, given the unfortunate economic timing.

The development of high quality data sets, for internet posting has not been successful to date. Serious efforts were made in two directions: an internet tomography data set, and a set of simultaneous tie series of router data. While both may still actually be posted, progress has been disappointingly slow. In retrospect, a central problem is that the “owners” of both projects were volunteer faculty (unsupported by SAMSI), which seemed to make sense at the time, as they were the most interested people. However, things moved too slowly, other interests got in the way, and momentum was clearly lost. A clear lesson is that such projects require an “owner” who has serious obligation to SAMSI.

# 2 Program on Data Mining and Machine Learning (DMML)

## 2.1 Introduction and Overview

Data mining and machine learning—the discovery of patterns, information and knowledge in what are almost always large, complex (and often unstructured) data sets—have seen a proliferation of techniques over the past several years. Yet, there remains incomplete understanding of fundamental statistical and computational issues in data mining, machine learning and large (sample size $n$ or dimension $p$) data sets.

The high-level goals of the SAMSI DMML Program were to advance this understanding significantly, 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.

**Program Leadership.** The Scientific Committee for the program consists of David Banks (Duke; Co-chair), Mary Ellen Bock (Purdue; NAC Liaison), Jerome Friedman (Stanford), Alan F. Karr (NISS; Chair), 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).

## 2.3 Working Groups

Scientific activities of the program occurred primarily in four working groups, which had distinct but overlapping foci. In particular, 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 Chair (Karr) and Co-Chair (Banks) with the four working group leaders.

The working groups, their leaders and participants, and general foci were:

**Bioinformatics**, 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.

**General Motors (GM) Data Analyses**, 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.)

**Large $p$, Small $n$ Inference**, 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), Jenhwa 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.

**Support Vector Machines**, 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.

**Theory and Methods**, 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), 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.

Detail on the goals and activities of the working groups follows in §2.3.1–2.3.5. 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.

### 2.3.1 Bioinformatics

**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.** Initial 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 include securing data sets, securing small company collaborators and papers submitted.

**Achievements.** To date, these include:

- 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.

### 2.3.2 GM Data

**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 in §2.1, 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.

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

### 2.3.3 Large $p$, Small $n$ Inference

**Goals.**[2] A stunning success would 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,

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

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

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), which may grow into papers, possibly a chapter in the monograph (§2.7).
- 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) will lead 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.

### 2.3.4 Support Vector Machines

**Goals.** Development of workable, interpretable multi-category SVM was identified as a particular notable success.

**Research Agenda.** This working group identified a series of "clusters" of interest, 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 $p$) 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 has been 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. <athematical 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. Once the computations are done, the cluster will add such theoretical justifications and submit the paper for publication.

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 being investigated, as are generalizations of the circle of ideas to nonlinear SVM.

**Multi-category SVM:** Ongoing investigation SVM methods for multicategory classification problems, including improvements of proximal SVM. Applications 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 is available and 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.

### 2.3.5 Theory and Methods

**Goals.**[3] A stunning success would 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.
- 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 will be completed by the end of 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 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.

## 2.4 Workshops and Course

Seven workshops were held as part of the DMML program, together with one course.

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

### 2.4.1 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,** on September 7–9, 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.

### 2.4.2 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.

### 2.4.3 Closing Workshop

The closing workshop for the program is scheduled for May 17–18, 2004. It will serve 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.

Approximately 50 attendees are anticipated.

### 2.4.4 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.

### 2.4.5 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.

## 2.5 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.[4]

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.

## 2.6 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.

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

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

**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.

## 2.7 Planned Follow-On Activities

**Proposals.** The Theory and Methods and Bioinformatics working groups have produced two proposals in the area of data analysis in metabolomics. One was submitted in March, 2004 to the National Institutes of Health (NIH), and the other will be submitted in April, 2004 to the Advanced Technology Program (ATP) program at the National Institute of Standards and Technology (NIST).

Discussions with GM regarding a NISS-led follow-on project will begin in April of 2004.

Other proposals are expected to be generated during the remainder of 2004 and beyond.

**Monograph.** A monograph on data mining will be written that pulls together research from all four working groups, which is planned to be submitted to the *ASA/SIAM Series on Statistics and Applied Probability*. David Banks and Alan Karr will be its editors. Approximately ten individual papers are anticipated, for which most commitments from authors are already in place.

**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,[5] 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 and will be 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: Ernest Fokoué, David Banks, Ashish Sanil
- International Society for Bayesian Analysis (ISBA): Bertrand Clarke, Feng Liang, Merlise Clyde, Susie Bayarri

[5] 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.

- 2004 Spring Research Conference on Statistics in Industry and Technology: Feng Liang, Jen-hwa Chu, David Banks
- International Federation of Classification Societies: Helen Zhang, Xiaodong Lin, Leanna House, Stanley Young, Jacqueline Hughes–Oliver
- 2004 Joint Statistics Meetings: David Banks; also, a topics contributed session organized by Murali Haran with Bertrand Clarke, Ernest Fokoué, Leanna House, and Katja Remlinger.
- COMPSTAT'04: David Banks
- Fields Institute Workshop (NPCDS—Canada): David Banks
- International Conference on the Future of Statistical Theory, Hyderabad (India): David Banks, Prem Goel, Murali Haran.

**Workshops.** A metabolomics workshop at NISS is planned for the fall of 2004.

**Education and Outreach.** A short course on data mining is being developed for JSM 2004 and the Hyderabad conference, in honor of C.R. Rao.

## 2.8 Anticipated Outcomes and Measures of Success

A complete evaluation of the DMML program will be included in the 2005 SAMSI Annual Report. At this point, anticipated outcomes and measures of success have been identified at both the program and working group levels.

### 2.8.1 Program Level

Outcomes at the program level are expected to be:

- 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.

### 2.8.2 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.

Evaluation at the working group level will use both these measures and the program-level measures.

# 3 Multiscale Model Development and Control Design

## 3.1 Introduction and Overview

The development of novel materials and transducer designs for high performance applications naturally leads to multiple scales in space and time which must be spanned to achieve the full capabilities of the compounds. Moreover, model development, numerical approximation, control design, and experimental validation and implementation involve both deterministic and stochastic components which must be addressed in concert to provide both fundamental understanding of the materials and to achieve stringent design and control specifications. The goal of the SAMSI Program on Multiscale Model Development and Control Design is to identify short and long term research directions deemed necessary to achieve both fundamental and technological advances in this rapidly growing field and to initiate collaborative research programs — between mathematicians, statisticians, materials scientists, engineers, and physicists — focused on the synergistic deterministic/stochastic analysis of advanced materials.

A number of these issues can be illustrated in the context of two prototypical materials, piezoceramics and ionic polymers, which are being considered for applications ranging from quantum storage to artificial muscle design. In both cases, the unique transducer properties provided by the compounds are inherently coupled to highly nonlinear dynamics which must be accommodated in material characterization, numerical approximation, device design and control implementation. For piezoceramic materials, deterministic energy relations can be constructed at microscopic or mesoscopic levels to quantify constitutive nonlinearities. To construct macroscopic models suitable for transducer or control design, however, variability in material properties, grain orientations, and field-stress distributions must be accommodated through either stochastic or deterministic homogenization techniques. The situation for ionic polymers is significantly more complex due to inherent coupling between chemical, electrical and mechanical properties of the compounds. Present investigations are focused on the use of Monte Carlo simulations to characterize material capabilities, and the development of energy relations through deterministic or statistical mechanics tenets is in its infancy. Furthermore, whereas statistical inference principles have been employed to construct multiscale hierarchical models for other physical phenomena, including biological and environmental mechanisms, application of these techniques to material characterization is in its nascency.

The real-time approximation of comprehensive material models for device design and model-based control implementation is a significant challenge which must be addressed before novel material constructs and device architectures can achieve their full potential. The realization of these goals requires the development of reduced-order models which retain fundamental physics but are sufficiently low-order to permit real-time implementation. One technique under consideration is based on Proper Orthogonal Decomposition (POD) techniques — also known as Karhunen–Loève or principal component analysis — which combines both statistical and deterministic aspects. Significant theoretical and implementational issues remain in the area of reduced-order model development for characterization and control design and this constitutes one of the focus topics for the

program. Furthermore, issues associated with reduced-order model development transcend this program and advances made here have the potential for making significant contributions to fields ranging from flow control to economic analysis.

The final component of the program focuses on control design, a crucial aspect of which is robustness with regard to disturbance and unmodeled dynamics. Deterministic robust control designs often provide uncertainty bounds which are overly conservative and hence provide limited control authority. Alternatively, one can provide statistical bounds on uncertainties. In the context of material models, one approach is to quantify the manner through which uncertainty is propagated between scales to address the goal of improving the characterization of uncertainty in system-level models used to quantify high performance transducers. To achieve the efficiency necessary for real-time implementation of model-based control techniques, one must typically employ reduced-order numerical models such as the previously described POD expansions. Hence model development, numerical approximation, control design, and experimental implementation must be considered in concert to achieve the unique capabilities provided by novel materials and transducer architectures.

Despite the highly interdisciplinary nature of the field, these components have primarily been investigated in isolation within both the mathematics and statistics communities. Even within the mathematics community, investigations have tended to treat model development, numerical approximation, and control design as segregated topics rather than synergistic facets of a unified field. However, to achieve the program objectives, it is necessary to investigate materials science, mathematical and statistical issues in concert, and this is the mandate of the SAMSI Program on Multiscale Model Development and Control Design.

The program includes the following components and activities:

- **Opening Workshop** (January 17–20, 2004): Leading researchers from a variety of fields encompassing multiscale analysis were asked to provide overview presentations and make recommendations concerning research directions deemed necessary to make significant advances in the field. This provided a framework for the research directions pursued in the program.

- **Workshop on Multi-scale Challenges in Soft Matter Materials** (February 15–17, 2004): Chair: Greg Forest (UNC). This workshop addresses the identification of open research questions and determination of collaborations and research directions required to provide significant advances in the fundamental understanding and utilization of novel soft matter materials — e.g., liquid crystal polymers.

- **Workshop on Fluctuations and Continuum Equations in Granular Flow** (April 15–17, 2004): Chair: David Schaeffer (Duke). This exploratory workshop will delineate research directions focused on model development, analysis, and numerical approximation techniques for granular flows.

- **Distinguished Lecture:** Professor Jonathan Chapman, Oxford University, March 2, 2004, “A Hierarchy of Models for Type-II Superconductors.”

- **SAMSI University Fellow:** Professor Arthur Krener, University of California, Davis.
- **SAMSI Postdoctoral Fellow:** Emily Lada.
- **SAMSI New Research Fellow:** Andrew Newell, University of California, Santa Barbara and North Carolina State University.
- **SAMSI Graduate Fellows:** Laura Ellwein (NCSU), Jon Ernstberger (NCSU), Joe Lucas (Duke), Eric Vance (Duke); their experience is discussed in Section I.C.
- **Short and Long-Term Visitors:**
- **Working Groups:** Three formal working groups have been formed to organize, pursue, and communicate research investigated during the program.

  – *Paradigms for Bridging Scales:* Led by Alan Gelfand (Duke) and Ralph Smith (NCSU)

  – *Control Design:* Led by Arthur Krener (SAMSI University Fellow from UC Davis)

  – *Homogenization:* Led by H.T. Banks (NCSU) and D. Cioranescu (U. Paris VI)
- **Courses:** Two courses are being taught in conjunction with the program.

  – *Mathematical and Statistical Techniques for Advanced Materials Synthesis*, Cross-listed at NCSU, UNC and Duke and co-taught by Ralph Smith (NCSU) and Alan Gelfand (Duke).

  – *Nonlinear Dynamics and Control*, Taught at NCSU by Arthur Krener (SAMSI University Fellow from UC Davis)
- **Technical Reports:** Technical reports produced or in progress are listed in Section I.G.

Details regarding these activities are provided in subsequent sections.

## 3.2 Program Objectives, Anticipated Outcomes and Recruitment

### 3.2.1 Objectives

The following program objectives were identified during the Opening Workshop. Details regarding these activities are provided in Section **??** in the context of the program working groups.

- Provide modeling frameworks for advanced materials, including piezoceramic, magnetic and shape memory compounds, liquid crystal polymers and granular compounds which, when appropriate, span spatial scales from angstrom-level quantum behavior to meter-level system dynamics. Temporal scales range from nanosecond to timescales as long as years. These frameworks will combine deterministic energy analysis with stochastic models and homogenization techniques to effectively bridge scales.

- Develop full-order numerical approximation techniques for high fidelity material characterization and reduced-order numerical models for real-time physical implementation. This will require a synergistic analysis of deterministic approximation issues in combination with stochastic frameworks such as those arising in Proper Orthogonal Decomposition (POD) techniques — also known as Karhunen–Loève or principal component analysis.

- Develop linear and nonlinear control frameworks which incorporate the manner through which uncertainty is transmitted across spatial and temporal scales to enhance efficiency and robustness. Significant emphasis will focus on the development of nonlinear control techniques that are feasible for real-time implementation.

- Construct experimental databases and perform validation experiments to establish properties and limitations of models and control designs.

- Catalyze collaboration between mathematicians, statisticians, material scientists, physicists and engineers to pursue the research program.

- Initiate research and educational programs necessary to achieve the long-term goal of providing deterministic/stochastic multiscale frameworks for the characterization, control, and *design* of novel compounds for aerospace, aeronautic, industrial and biomedical applications. Communicate with program managers and funding agencies to ensure that they know the directions projected by the program.

- Provide a nucleus for developing models, numerical techniques, and control theory through workshops, research visits, colloquia, classes, and maintenance of the program web site `http://www.samsi.info/200304/multi/multi-home.htm`.

### 3.2.2 Outcomes

Expected outcomes at the program level are the following. The methods used to document each outcome are indicated in brackets. A number of the outcomes will still be in progress at the program conclusion and post-program evaluation will be be performed in these cases.

- Significant research accomplishments by the working groups leading to collaborations and interdisciplinary research with the goal of starting to write papers and research proposals by the end of the program. *[Documentation: A fundamental result will consist of documents defining future research directions identified during the program. These documents will be disseminated to program managers at funding agencies and the general research community. We will also document the number of papers, proposals, and research visitors for the program.]*

- Significant educational benefits for mathematics and statistics graduate students, postdocs, and faculty through the two semester-long courses. This will provide a common theoretical basis and vocabulary to pursue truly integrated research focused on the deterministic/stochastic multiscale analysis of advanced material systems. *[Documentation: Formal evaluations and feedback from course participants.]*

- Significant educational experience for SAMSI graduate students and postdocs as well as visiting students and postdocs. This will strongly enhance their perspectives regarding future research directions considered as important by experts in the field. *[Documentation: Formal evaluations and feedback from participants.]*

- Provide a focal point within the mathematics and statistics communities for synergistic multiscale analysis of advanced materials through the visitor program and program website. *[Documentation: Feedback from visitors.]*

- Dissemination of research directions and results through presentations at international conferences (e.g., Joint Statistical Meetings (JSM) in Toronto, August 8-12, 2004), articles in society publications (e.g., SIAM News), and publication of the recommendations from the Soft Matter Workshop in an appropriate proceedings volume. *[Documentation: A record of all presentations and publications will be maintained.]*

## 3.3 Program Activities

### 3.3.1 Program Leadership

The scientific committee consisting of Ralph C. Smith (Chair: NCSU), Alan Gelfand (Co-Chair: Duke), Doina Cioranescu (Universite Pierre et Marie Curie), Greg Forest (UNC), Murti Salapaka (ISU), David Schaeffer (Duke), Chris Wikle (Missouri) and Margaret Wright (NYU) coordinate and guide the program activities.

### 3.3.2 Opening Workshop

The SAMSI Opening Workshop for the Program on Multiscale Model Development and Control Design was held on January 17–20, 2004. The goal of the workshop was to initiate discussion focused on identifying avenues of multiscale analysis crucial for the success of advanced materials in present and projected transducer designs for fields ranging from quantum computing and nanopositioning to liquid crystals and granular flows. A prominent theme throughout the workshop was the necessity of exploiting the natural synergy between model development, numerical approximation, and control design utilizing a combination of deterministic and stochastic specifications to explain and achieve the novel performance capabilities of advanced materials.

Tutorials on *Energy Techniques for Multiscale Modeling*, *Principles of Multilevel Stochastic Modeling*, and *Control Design for Nonlinear Systems* were offered on the first day of the workshop

with 75 participants in attendance. The program during the remaining three days consisted of six focus sessions on topics pertaining to multiscale materials analysis, oral poster previews, and a poster session. Each session consisted of two 40 minute overview presentations followed by 40 minutes of discussion focused on identifying directions necessary to achieve significant scientific advances in that facet of the field. The open forums were expanded in the final session of the workshop to identify objectives and research directions to be pursued in the program.

A total of 92 individuals registered for the workshop of whom 40 were affiliated with mathematics departments, 24 were in statistics departments, and 28 had affiliations in physics, materials science or engineering. The demographics for the second day are representative of those for the full workshop and can be summarized as follows. There were 82 attendees of whom 19 were women, 4 were African American and 1 was Hispanic. The total attendance for each day is summarized in the Appendix.

During the open forum on the final day of the workshop, three focus areas were identified and suggested as topics for working groups: (i) Paradigms for bridging scales, (ii) Control design, and (iii) Homogenization. The first focuses on the synergistic construction of deterministic and hierarchical models based on inference analysis for a prototypical material. It was suggested that the control group focus on the development of robust control techniques. To permit real-time implementation, the development of reduced-order methods comprises one aspect of the control design. Finally, it was recommended that the homogenization group focus on deterministic and stochastic homogenization techniques feasible for numerical and eventual experimental implementation. It was stressed during the forum that groups communicate frequently through joint meetings and overlapping participants and this has been promoted during the program.

### 3.3.3 Workshop on Multi-scale Challenges in Soft Matter Materials (Greg Forest, Organizer)

At the outset, the stated primary goal of the meeting was to identify and develop collaborations across disciplines and across scientific method of emphasis. There is a clear advantage to merge theory, modeling, simulation, and experiment toward deeper understanding of the many mysteries in soft matter. The mathematical community is warming up to the reality that soft matter systems are pervasive across biology, natural physical systems, and technology, and yet we do not have robust model equations, primarily because the "noncovalent" physics and chemistry that dominates such forms of matter are only partially understood. The numerical algorithms for the simplest classes of isotropic, continuum models of viscoelasticity simply fail in the limit of high Deborah number when elastic timescales are shorter than fluid timescales.

This unsettled ground for mathematics and theory is ripe for statistical intervention. Proposed models are routinely generated and tested against sparse data collected in experiments, with no overarching application of model validation principles across model space, and very little mention of statistical design of experiments. The data collected, either from imaging or force detection instruments, is rarely directly related to model variables; thus, interpretation of data is a major issue of unknown proportion, and data mining is rarely discussed in this context. Often, measurements

are made of moments of distribution functions, e.g., degrees of entanglement or degrees of coil versus stretch in long-chain molecules, or orientational distributions of anisotropic molecules. These issues are central to the classical problems of non-uniqueness of moment truncations moment-closure models.

The undercurrent of multiple scales in space and time is now evident and acknowledged, further compelling the need to intelligently blend all components of the scientific method toward unraveling the phenomena, form, and function of soft matter. The mix of natural, biological, and synthetic material systems that was represented at the meeting helped all participants to further reach for unifying perspectives, as well as address focused model systems. What classification of soft matter model systems will emerge to replace the classical continuum mechanical, coarse-grained equations and model systems of mathematical physics? This and many other meta-science questions are the topics of workshops and conferences across the science, health, and technology communities, and are becoming necessary topics of debate as the cultures of biology and medicine merge and interact with those of physical, mathematical, statistical, and computational science. . Increased awareness is mandated for any discipline and culture to access and compete in major funding initiatives of the federal government (e.g., the multi-agency Nanoscience and Technology initiative and the NIH Roadmap).

The organizing committee was chaired by Greg Forest, and assembled to reflect all components of the modern scientific method: experiments, theory, modeling, and computation. Day 1 focused on overview presentations by the Organizing Committee, followed by a rousing poster session which commanded both floors of the NISS building. Day 2 then went into more detail with specific short talks in each methodology, again with the presentations focused on current advances and open questions on the near horizon. Day 3 was dedicated to synthesis, and an attempt to put down in writing the collaborative projects that were spawned during the meeting. A brief summary of Days 1, 2, and 3 follow.

**Day 1–Overview Presentations & Poster Session:** Day 1 began with an afternoon session. We started with a representative group of soft matter experimentalists who are solicitous of theory, modeling, and simulation in their worldview: Patrick Mather, a materials chemical engineering from U. Connecticut, and Maria Kilfoil, a materials physicist from McGill University. Maria and Patrick presented an overview of physical phenomena and principles of soft matter. Maria focused on soft solids and the open questions surrounding energetic scaling laws and equations of state, from atomic scales to colloids and granular systems. Maria further explored the use of experimental tools, such as confocal microscopy, to predict equations of state to resolution comparable to the transparency of simulations. Patrick addressed why it is worth building theory-computation-experimental relationships and teams. He enticed the audience with experiments of durotaxis (sensing interface compliance) of living cells on soft materials, and a liquid crystalline elastomer swimmer. "Model that!" was the challenge. Maria and Patrick put together a Day 2 session of experimentalists and phenomena that exposed problems and challenges for the mathematics and statistics community.

The second overview was given by the computational science sub-committee of Michael

Graham and Yannis Kevrekidis. Because of the engineering, technological and health importance of soft matter, from the foods industry to personal care products to drug delivery to traditional polymers to nano-composite materials to biological fluids, numerical technology has been advanced to guide and sometimes supplant expensive experimentation and prototyping of engineering design. This has occurred by adaptation of many numerical advances of the applied mathematics community, but it is fair to say that the heavy lifting and innovation in scientific computation approaches to soft matter have not been dominated by applied mathematicians or statisticians. Mike gave a compelling and informative overview of computational challenges in the flow of soft materials. An essential feature of soft matter is the prevalence and critical role of nontrivial microstructure. All properties are largely determined or modulated by non-covalent effects: electrostatic, hydrodynamic, excluded volume, slippage or adhesion of molecules, etc. The viscoelastic analog of Reynolds number is the Deborah number, describing the ratio of liquid and material timescales, and the rubber meets the road (actually, the fluid) when the Deborah number is O(1) or larger. Then stresses are stored rather than immediately released through flow. How those stresses are parsed and the timescales upon which they are stored and released is a central focus of computation and modeling approaches. The descriptive variables and numerical methods were described, comparisons of molecular and continuum scale models and methods, new issues and approaches for macromolecules in confined geometries (e.g., for DNA microfluidic devices), Lagrangian versus Eulerian methods, many open issues in the dynamics of flowing suspensions (shear-induced migration and diffusion, structure formation, jamming, statistics of velocity fluctuations), multi-component systems and mixing versus de-mixing (phase separation), parameter estimation methods for reduced models and the bridge to statistical tools and methods, sensitivity analysis of stochastic simulation methods, were all highlighted. Yannis gave an overview of what one might call multiscale meta-computation principles. He discussed when and how one might use stochastic, detailed microsimulators or black box codes to do equation-free computation. A central idea is to finesse the derivatives (Jacobian) of the equation you don't have by short-time runs, and then use it to do Newton-Raphson iteration and find bifurcations. The idea of Herb Keller, to use a legacy code to find and compute the subspace of slow dynamics, and then do fixed point analysis from the dynamics of this slow manifold, was lucidly presented.

The final overview was given by the theory and modeling sub-committee of Qi Wang, Alejandro Rey, and Michael Rubinstein (who did not speak, yet made a big impact on the choice of participants). It is both exhilarating and intimidating for theorists and modelers to work in a field where the fundamental, first principles-based, equations are not known. Indeed, the correct primitive variables are not always clear: position, velocity, deformation, stress, orientation, conformation, etc. It is most often the case that posited model equations, guided and conditioned by continuum mechanical invariance, symmetry, and entropy inequalities, are the rule of order. This is to be contrasted with models for aerodynamics, meteorology, or turbulence, where the Navier-Stokes equations are more or less accepted, and the challenge is to derive accessible (in theory or computation) models from the primitive equations.

All forms of soft matterĚ.from polymers to bio-fluids to nano-composites to colloids and suspensions to granular materials, from bulk systems to confinement and interfacial behaviorĚ.remain immature with respect to fundamental understanding. This state of affairs is very poorly advertised in the mathematics and statistics cultures, and only now taking a prominent position in the headlines of the physics and engineering cultures as the biological, medical, and technological importance of soft matter systems become apparent. Qi Wang gave a comprehensive overview of theory and modeling of soft matter materials, which have both fluid-like and solid-like properties, from molecular-based models to continuum mechanical theories, from isotropic to anisotropic polymeric systems. The lecture provided a context for participants to place the specific models used by others at the meeting. Alex Rey gave a penetrating insight into interfacial phenomena in soft matter systems. Here there is very little known about the basic principles of how to describe the interface where soft matter meets more mundane matter such as solids or liquids or gases. Alex motivated the theory with physical phenomena from experiments and Nature, and then laid down the early stages of fundamental theory and principles.

From Wang & Rey's lectures, one could gain an appreciation for why soft matter is so nebulous in terms of first principles descriptionĚ.such materials are defined in terms of what they are not (simple viscous fluids or simple elastic solids). So, until one knows precisely what soft matter is, and how such materials inherit lengthscales and timescales due to short-range and long-range interactions of attractive and repulsive type, the quest for precise mathematical models of equilibrium and non-equilibrium behavior will remain wide open. Soft matter deserves to be a poster child for multi-scale in time and space, with the scales and their genesis yet to be discovered.

After the three overview presentations, the atmosphere and participants were ripe for cross-linking! Since the organizers had communicated numerous times through email, the overviews of each "method sub-committee" had frequent contacts with one another. We then transported over to the NISS building for a truly engaging poster session. All participants were repeatedly hammered with the notion that the poster session was the centerpiece of the meeting. We left 3 hours on the schedule for heavy hors d'ouevres, refreshments, and posters by all. This event was aimed at crossing language barriers and allowing everyone a chance to meet one another and view a condensed version of their research emphasis. This was especially important for a short Workshop with participants who have never assembled together before, and which did not allow for everyone to give an oral presentation. By all accounts, the only negative aspect of the poster session was we needed more time to view so many cool posters!

**Day 2–Special Sessions & Invited Talks:** Day 2 consisted of an artfully chosen set of 20 minute presentations with 10 minute discussion. The talks represented a diverse set of important advances in theory, computation, modeling, and experimental phenomena and technique. We also had a healthy mix of junior scientists, and it so happened that women in science were equal if not dominant at this workshop. The list of talks are given in the Appendix.

**Day 3–Toward Outcomes of the Workshop:** An advertised emphasis of the workshop was to actively identify and define collaborations. Our primary goal was to bring diverse scientists together who have strong potential to build interdisciplinary teams. This goal was continually reiterated leading up to and during the meeting. After a relaxing Day 2 evening with a dinner outing in Chapel Hill, the morning of Day 3 was dedicated to synthesis: we solicited the participants to brainstorm collectively for about an hour, and then to form breakout sessions to write down the collaborative projects spawned during the meeting. After an hour, we reconvened for the last hour to present the collaborative project lists and people, and to suggest ways that SAMSI might facilitate these collaborations.

We close the Report with these two products of the meeting.

**SAMSI Facilitation Suggestions:** The participants suggested several mechanisms by which SAMSI can facilitate continuation of the collaborations and contacts that were made during the Workshop, and mechanisms for dissemination of outcomes and broader suggestions to the scientific community.

1. A website for the workshop that will remain linked from the SAMSI webpage for ease of reference of people and topics, and for archival purposes. The webpage will include: workshop contributions (poster files or scanned versions of poster content; lecture files or scanned versions); webpage links for everyone who requests such a link; a list of participants, including their affiliations and email addresses; a list of the working groups proposed and listed in this report; and the Final Workshop Report. Subsequent products of these collaborations, in the form of published papers and even research proposals, could be subsequently linked to the webpage.

2. A distribution list of participants can be maintained and available upon request by mail. An email listserve was not recommended due to potential for SPAM abuse.

3. A Virtual Center or Virtual Collaborative Group concept was raised by Patrick Mather. The idea is to form virtual centers around the collaborations that were spawned at the meeting, and their evolution into possibly larger collaborative efforts. The synergies of working groups from various disciplines, and from the mixture of theory, modeling, simulation and experimental emphases, should be presented to funding agencies to underscore the leveraging and value added by these working groups, at no additional cost to each program.

4. Strong consideration for follow-up workshops as the collaborations and suggestions and outcomes of this workshop come into fruition.

5. Some form of "Proceedings" was discussed. A better phrase is needed perhaps, but the organizing committee and participants agreed to debate on the proper format and vehicle for documenting the workshop activities and contributions, and for providing a roadmap of future challenges and directions ripe for interdisciplinary collaboration that junior scientists might be guided by.

6. Peter Constantin raised the contrast between European models for supporting and promoting broad-scale research programs on timely topics, and the lack of such mechanisms in the U.S. The important fundamental and applied aspects of soft matter materials and phenomena compel a conversation with leaders in the academic community, NSF, DoD and DoE, and possibly DHS. How might the participants forge a conversation and possible emphasis on support of research collaborations, such as those we have identified herein, in the emergent area of soft matter? Subsequent to the meeting, Eitan Tadmor of the University of Maryland communicated that such conversations are underway at NSF to support international virtual centers around emerging topics ripe for mathematics and statistics.

7. Another challenge is how to infuse statisticians into our discussions and projects. Leon Glaser and Chuanshu Ji attended many, if not all, sessions. It would be beneficial for all if a concerted effort were made to engage statisticians in the proposed working groups.

**A Portfolio of Collaborative Projects, Themes, and Participating Scientists:** On the last morning of the workshop, the following projects & collaborators were identified.

1. Hydrodynamics of micellar solutions; Amy Shen and Pam Cook

2. Relationships between granular and polymer chains; Amy Shen, Bo Li

3. Models of snail locomotion; Linda Smolka, Peko Hosoi

4. Connections between mucus rheology and its role in transport across biological systems; Greg Forest, Peko Hosoi, Mike Graham, Rich Superfine (leader of the UNC Virtual Lung Project), Pat Mather, Michael Rubinstein

5. Two-dimensional polymers; ZhiFeng Huang, Shenda Baker, Peko Hosoi

6. "Pearling Threads" in suspensions; Peter Mucha, Peko Hosoi

7. Test problems for quantitative testing of sedimentation codes; Simon Tavener, Peter Mucha [bifurcation at Re=0; error introduced by operator splitting; error introduced by approximation for multiple particles; error introduced by multiscale approximation; comparison with lattice Boltzmann simulations]

8. "Molecular particles": 3d spherical dendrimers, generation 3 layers; transition from generation 5 to 6 (a few nm) as fluid behavior stops yet the drop doesn't spread! Generation 6 or higher has 2 glass transitions; soft-core potentials (granular is hard core); static friction-like behavior; gravity doesn't matter; Karen Daniels, Sergey Sheiko, Andrey Dobrynin.

9. Complex fluids in confined geometries. Shear-induced migration-competition with adsorption; flow of confined suspensions, jamming, structure formation; Mike Graham, Pat Mather. This project has natural ties to many participants who may want to join: Karen Daniels, Maria Kilfoil, Bob Behringer, Greg Forest, Qi Wang, Ruhai Zhou, Hong Zhou, etc.

10. Liquid crystal optics: theory, modeling and computation, homogeneous and inhomogeneous media. Chuck Gartland, E. McKay Hyde, Alex Rey. Natural ties to Forest-Wang-Zhou-Zhou group in nematic polymer film optics.

11. Effective thermal conductivity properties and anisotropy of polymer nano-composites. Steve Rosencrans, Xuefeng Wang, Bill Mullins, Eric Choate, Greg Forest, Rob Lipton. Given the orientational molecular distribution function in space and time for flow-processed, low volume fraction, spheroidal nano-inclusions, find the effective thermal conductivity tensor via homogenization, and then estimate and compute the long and short term behavior of heat flow and distribution due to applied thermal conditions on thin films. Strong ties to Pat Mather, Qi Wang, Ruhai Zhou, Hong Zhou, Xiaoyu Zheng, Leonid Berlyand on related, and even multifunctional, nanocomposite properties: percolation thresholds (Rob Lipton); effective mechanical, electrical (Xiaoyu Zheng), and permeability properties.

12. Rigorous mathematical theory of nematic polymers at rest and in linear flow fields. Peter Constantin, Edriss Titi, Yannis Kevrekidis, Ruhai Zhou, Qi Wang, Greg Forest

13. Macroscopic and mesoscopic consequences of microscopic symmetries. How can one get a macroscopic stress from microscopic properties, for example. Bulk properties from underlying symmetries. Leonid Berlyand, Peter Constantin, Maria Kilfoil. Related to methods for derivation of stress fields in mesophases of nematic polymers from microscopic details by Qi Wang.

14. Complex fluid interfaces. Contact line dynamics; interfacial dynamics under flow; coating flows and thin films; interfacial non-Boussinesq mechanics. Alex Rey and Mike Graham. Related to other projects on density variations in nematic polymer nano-composites, and on deformable solid membrane-complex fluid and gel interactions in biological systems (snail mucus of Peko Hosoi and virtual lung project of Superfine, Rubinstein, Forest)

15. Filtration for non-Newtonian fluids. Chris Cox, Peter Mucha

16. Mixing nanocomposite materials. Stephen Bechtel, Chris Cox

17. Graded index photonic crystals and band gap computation (made from selective etching of polymers); McKay Hyde, Maria Kilfoil

18. Wave ray tomography to infer contact forces in 3d granular materials using birefringence and fluorescence; Robert Behringer, Karen Daniels, McKay Hyde

19. Modeling of magnetorheological fluids; application to control. Qi Wang, Steve Bechtel, Mike Graham, Greg Forest

20. Droplet relaxation of liquid crystals for interfacial tension measurements; Pat Mather, Alex Rey

21. Isotropic-smectic transition; Pat Mather, Alex Rey

22. Nemato-electrowetting (long term project); Pat Mather, Alex Rey

23. "Yielding" of textured nematic elastomers; Pat Mather, Alex Rey, Qi Wang

24. Asorption of polyelectrolytes; Pat Mather, Andrey Dobrynin

25. Patterns that exist in isotropic-nematic transition, with existing data; Pat Mather, Greg Forest, Qi Wang

26. Trapped entanglements in liquid crystal elastomers; Pat Mather, Michael Rubinstein

27. Deposition and transport of aerosols, bacteria, and other pathogenic invaders in pulmonary pathways; Simona Mancini-Cordier, Greg Forest, Rich Superfine and the Virtual Lung Project at UNC.

It is anticipated that the participants will remain in contact on these diverse topics, and possibly meet again at the IMA in Minnesota in Fall, 2004, when the IMA will launch a special year on soft matter and related topics. Several key participants from this workshop were suggested to the IMA organizing committee for short term visits and invited lectures.

### 3.3.4 Workshop on Fluctuations and Continuum Equations in Granular Flow (David Schaeffer and Robert Behringer, Organizers)

The meeting was a great success scientifically. It brought together people of different backgrounds–physicists, civil engineers, and mathematicians.

The main statistical issue in granular materials is to derive continuum constitutive laws from micromechanical information, which will be of a statistical character. In three of the five sessions, this issue was addressed from several disciplines–experiments, physical theory, and mathematical analysis. This problem is far from solved, and there was a stimulating variety of viewpoints.

The applications session focused on the flow of granular materials through a silo. Here the main issue is reliability. The behavior of flowing and stagnant granular materials is very counterintuitive in many aspects. Extreme fluctuations from the average behavior happen only too often, and these fluctuations are typically neglected in design. (Moreover, because of our inability to describe such materials adequately, even an enlightened designer does not know how to allow for them.) Consequently, silo failures are painfully common–100 to 1000 times more frequent than failure of other engineering structures. Despite the frequency of collapses, ordinary statistical analyses are not possible because of lack of data–instrumentation to collect the relevant data is extremely expensive, and moreover most designers are unaware of what information is needed.

The fifth session introduced three related classes of problems. (i) Phase transitions in ferromagnets: In such systems, the statistics of spin flips depend crucially on an order parameter governing randomness in the system. It seems likely that the ferromagnetic system may provide insights

to physical avalanches and landslides; indeed the term “avalanche" is used to describe events in which a large number of spins flip simultaneously. (ii) Underwater landslides: Experiments on such phenomena were presented. In this case the interaction of the granular material with the fluid surrounding it leads to new, interesting phenomena. (iii) Glasses: Spin glasses are a disordered system much studied in the physics community. It is hoped that this subject may provide guidance on how to approach a statistical treatment of granular materials.

In the early afternoon of both days we divided into discussion groups. The groups were allowed to form spontaneously, and each time there were about four groups. These groups were an important, stimulating part of the meeting. One group focused on the extent to which the response of a granular material can be described by nonlinear elasticity. A particular problem deserves special focus: i.e., how a delta-function addition to the stress at the surface of a granular material propagates into the interior. Of course, because granular materials are disordered, this question must be addressed statistically, both experimentally and theoretically. Some theories and some experiments suggest that the distribution of added stress at depth will have a two-peaked distribution, and it is argued that such behavior provides evidence that the equilibrium PDE describing a granular medium are hyperbolic. However, as was discussed at the session, nonlinear elasticity can also lead to such a response. We also discussed how the elastic behavior of a medium can be derived from micromechanics.

A variety of state variables have been proposed to characterize granular materials and construct constitutive equations: (i) a scalar variable is at the basis of Aronson’s Ginzburg-Landau type approach; (ii) a tensorial state variable emerges in STZ theory; several definitions of granular temperature have been proposed in relation to fluctuations of either (iii) kinetic, or (iv) elastic energy–and I would add (v) enthalpy; (vi) tensorial kinetic temperatures have been introduced to account for directional fluctuations of velocities. We need to clarify the relevance and relation between these different variables: (1) do we need tensorial state variables, and if yes, do we need tensorial temperatures? (2) What is the relevance of kinetic vs elastic energy fluctuations? (3) Since deformation is associated to changes in the density, how should this be taken into account?

## 3.4 SAMSI University Fellow

**Arthur Krener** (01/01 – 5/01): Arthur Krener received his PhD in mathematics from the University of California, Berkeley, in 1971. Since 1971 he has been at the University of California, Davis where he has been Professor of Mathematics since 1980. In 2002 he was named a Distinguished Professor of Mathematics. He has held visiting positions at Harvard University, the University of Rome, Imperial College of Science and Technology, NASA Ames Research Center, the University of California, Berkeley, the University of Paris IX, the University of Maryland, the University of Newcastle, Australia and the University of Padua.

Professor Krener is a Fellow of the IEEE. His 1981 IEEE Transactions on Automatic Control paper with Isidori, Gori-Giorgi and Monaco won a Best Paper Award. His 1977 IEEE Transactions on Automatic Control paper with Hermann was recently chosen as one of 25 Seminal Papers in Control in the last century by the IEEE Control Systems Society. He was a Fellow of the John

Simon Guggenheim Foundation for 2001–2002.

His current research interests are in developing methods for the control and estimation of nonlinear dynamical systems and stochastic processes. Along with Wei Kang he developed the theory of control bifurcations. A control bifurcation is the loss of stabilizability of an equilibrium of a control system. Krener is also working on methods of developing low order models for control of complex systems. Recently he proved that the extended Kalman filter is a locally convergent observer for a broad class of nonlinear systems.

## 3.5 SAMSI Postdoctoral Fellow

**Emily Lada** (01/01/04 – 8/15/05): Emily Lada received her BA in mathematics from the University of North Carolina at Chapel Hill. She also hold MS and PhD degrees in operations research from North Carolina State University. While at North Carolina State University, she was awarded a Selected Professions Fellowship from the American Association of University Women, a General Electric Faculty for the Future Teaching Fellowship, and a North Carolina State University Dean of Engineering Fellowship. Before coming to SAMSI, Emily was a Senior Research Scientist at Old Dominion University's Virginia Modeling, Analysis, and Simulation Center (VMASC) where she participated in the validation of a military logistics model designed to calculate the feasibility of transportation and sustainment of military forces. She also assisted in the development of material for overview courses in modeling and simulation.

Emily's main research interests are in the area of large-scale simulation modeling and analysis. She is currently investigating the development of a simulation model of nafion, which is one of the ionic polymers presently under investigation for use in applications ranging from artificial muscle design to chemical sensing. The polymer chains making up nafion are composed of a hydrophobic back bone with hydrophilic side chains. In order to develop an energetics model for the hydrophilic regions, it is necessary to estimate the material properties of the hydrophobic region. The bulk material properties of the hydrophobic region depend on how stretched out the chains are. This effect is modeled by tracking the distribution of end-to-end distance for a large number of chains.

## 3.6 SAMSI New Research Fellow

**Andrew Newell** (01/01 – 5/15): Andrew Newell received his PhD in geophysics at the University of Washington in 1997. He did brief postdoctoral stints at the University of Washington and Scripps Institution of Oceanography. From 2000 to 2003 he was a Research Assistant in the Department of Geological Sciences at the University of California Santa Barbara. He is presently a Research Associate in the Center for Research in Scientific Computation at North Carolina State University.

His research revolves around the magnetic properties of rocks with applications to paleomagnetism and environmental magnetism. Areas of concentration include micromagnetic modeling, Preisach theory, low temperature phase transitions in ferromagnets, and magnetotactic bacteria with applications to exobiology.

## 3.7 Short and Long-Term Visitors

**Short-Term Visitors:**

- **John Whiteman** and **Simon Shaw** (Senior Faculty – Brunel Unversity, UK, Applied Mathematics, November, 2003): Pursued research in collaboration with H.T. Banks, J. Hood, N. Medhin and G. Pinter on multiscale model development and inverse problems for polymers and filled elastomers in which molecular-based formulations with uncertainty — characterized by probability distributions that must be estimated — are constructed to explain macroscopic hysteretic phenomena.

- **Gabriella Pinter** (Junior Faculty – University of Wisconsin, Milwaukee, Applied Mathematics, November, 2003): See the previous research description.

- **Doina Cioranescu** (Senior Faculty – Universite' Pierre et Marie Curie, Applied Mathematics, 1/16 – 1/22): Presented an invited lecture during the Opening Workshop and collaborated with local scientists in the Homogenization working group.

- **Simona Mancini** (Postdoctoral Fellow – Universite' Pierre et Marie Curie, 2/14 – 2/18): Attended the Workshop on Multi-scale Challenges in Soft Matter Materials, participated in the working groups and attended the SAMSI class.

- **Jonathan Chapman** (Senior Faculty – Oxford University, Applied Mathematics, 2/29 – 3/4): Presented a Distinguished lecture, met with local and visiting scientists, and participated in the working groups.

**Long-Term Visitors:**

- **Leon Gleser** (Senior Faculty – University of Pittsburgh, Statistics, 1/23 – 2/27): Professor Gleser's current research interests are in linear and nonlinear measurement error regression models, theories of statistical inference, statistical meta-analysis, applications of statistical model and methods to the biological, physical and behavioral sciences.

  Professor Gleser participated in the working group on "Paradigms for Bridging Scales" and met with students, postdocs and both visiting and local faculty during his visit.

- **Nicholas Zabaras** (Senior Faculty – Cornell University, Applied Mathematics, 2/2 – 3/5): Professor Zabaras' research focuses on the development of robust multiscale material models which can be employed both for optimal material design and optimal design of systems exploiting advanced materials. This includes the development of spectral stochastic methods for modeling uncertainty propagation in the analysis, optimization and design of continuum systems, and the development of multi-length scale algorithms for the analysis and design of microstructures in engineering materials. Aspects of his research also focus on the development of real time feedback mechanisms for the control of complex material processes in the presence of uncertainty.

Professor Zabaras participated in working groups on “Paradigms for Bridging Scales” and “Control Design.” He also worked with students, postdocs and visiting and local faculty during his visit.

- **Ricardo del Rosario** (Junior Faculty – University of the Philippines, Applied Mathematics, 4/9 – 5/7): Professor del Rosario’s research focuses on the development of full and reduced-order numerical approximation techniques for smart material systems including piezoceramic shells. He also investigates the development of control techniques for smart material systems including full-state feedback designs and compensators employing asymptotically convergent state estimates.

  Professor del Rosario will participate in all three working groups, contribute to the SAMSI class, and pursue research with the SAMSI postdoc for the program as well as local and visiting scientists.

## 3.8 Working Groups

**Paradigms for Bridging Scales**

This group has identified the following high level objectives.

- Pick a prototypical or representative material and build, to the extend possible, a full multiscale analysis ranging from microscopic to system level. This will employ techniques including energy analysis, meta-modeling, data aggregation and homogenization theory. An important consideration when choosing the material is ensure that we have sufficient data at the different scales to motivate physical mechanisms and permit validation. At the microscopic and mesoscopic levels, this data may be simulated using high resolution codes.

- Employ inference analysis to develop hierarchical models for material characterization and control design, and jointly analyze hierarchical and deterministic models to construct unified characterization frameworks.

- Once the multiscale framework is understood for the prototype, we will begin to extend it to novel compounds projected for advanced applications — e.g., piezoceramics, magnetic and shape memory compounds, liquid crystal polymers.

- A fundamental component of the investigation will focus on the use of statistical methods and models to reduce computational times for 2-D and 3-D materials models. This is crucial both from the perspective of transducer design and the real-time implementation of linear and nonlinear control techniques. This component will take as a starting point the development of reduced-order numerical models using Proper Orthogonal Decomposition (POD) techniques.

- Identify a source of multiscale data to motivate and validate mathematical/statistical models. Initial data will come from NC State University and GA Tech.

- Determine a prototypical system for investigating failure analysis in high performance systems through characterization of distribution tails. One goal is to construct a framework which facilitates identification of microstructures given macroscopic measurement. To the extent possible, this will be investigated in collaboration with a modeling group from **Data Mining** which is presently in place.

**Control Design**

A high level goal of this group is to develop control theories and algorithms which employ multiscale analysis to enhance robustness and facilitate physical implementation for advanced material transducers operating in highly nonlinear and hysteretic regimes. To accomplish this, the following objectives have been identified.

- Control inputs and sensing often occur on multiple levels. Develop numerical techniques and homogenization theory to integrate multiscale inputs and sensing.
- Quantify the manner through which uncertainty at the fine scale is manifested at coarser scales to enhance the robustness of resulting control theories and designs.
- Develop stochastic robust control theories, designs and algorithms, based on the models developed under "Paradigms for Bridging Scales," which incorporate robustness by quantifying the manner through which uncertainty is propagated between scales.
- A crucial component of model-based control design entails real-time implementation. A significant objective of the group is to employ deterministic and statistical model reduction techniques to construct low-order models suitable for linear and nonlinear control implementation.
- Experimentally implement the robust nonlinear designs to demonstrate their feasibility for systems employing advanced materials.

**Homogenization**

Homogenization techniques provide a natural bridge between a number of stochastic and deterministic multiscale issues for advanced materials. A high level objective of this group is to start with the theoretical deterministic foundations discussed by D. Cioranescu during the opening workshop and investigate the manner though which this framework can be formulated and implemented for advanced materials. A second objective is the investigation of stochastic homogenization techniques to construct macroscopic material parameters from meso- or microscopic energy relations for the constituent materials.

# 4 Education and Outreach Program

The SAMSI Education and Outreach (E&O) program has continued to evolve and expand during the second year of SAMSI operation. In addition to a goal of major education and outreach to members of the scientific community at non-local (i.e., non-Research Triangle) institutions, the program now also focuses on significant participation by SAMSI Graduate Fellows and Postdocs as a means to train young investigators in the important tasks and methodologies of outreach and education. The response and performance of SAMSI Graduate Fellows and Postdocs in this effort has been enthusiastic and most effective.

The SAMSI Education and Outreach program for 2003-2004 included the following activities:

**Outreach 2-day workshops for undergraduates** were held on November 14-15, 2003 and February 13-14, 2004. Undergraduates and faculty mentors were invited to SAMSI for a day and half long program during which the members of the SAMSI Directorate gave an overview of SAMSI and current and future programs, and the opportunities for participation were discussed. Members of the Data Mining and Machine Learning (DMML) Program (led by Alan Karr) then gave a focused one day plus presentation on DMML with experimental data driven demonstrations.

**Undergraduate Interdisciplinary Workshop:** A one week workshop for approximately 15 undergraduates was held June 9-13, 2003 and is scheduled to be repeated May 30 -June 4, 2004 with a somewhat increased number of participants. During this week, the students are exposed to an intensive program involving formulation of inverse problems, hands-on collection of experimental data, and mathematical and statistical analysis of data. The problem chosen (vibrations of a cantilevered beam mounted with sensors and actuators) is a paradigm for the theme of SAMSI (an interdisciplinary approach involving applied mathematics, statistics, and applications in domain science to solve complex practical problems) and embodies the principles to which we are trying to attract young mathematicians/statisticians. All presentations and tutorial sessions are organized and given by SAMSI Graduate Fellows and Postdocs under close supervision of the SAMSI Associate Director for E&O.

**Industrial Mathematical and Statistical Modeling Workshop:** This ten day workshop for Graduate Students was held July 21-30, 2003. This workshop involved 37 graduate students selected from a competitive national pool along with representatives from six industrial/government labs to work in teams on problems posed by the industrial representatives. Team leaders and topics in 2003 were: David Dausch and Scott Goodwin, MCNC Research and Development Institute (Electrostatic Operation of a MEMS Flexible Film Actuator); William C. Hunter, Federal Reserve Bank of Chicago (Rational Price Limits/Circuit Breakers in Futures/Equity Markets); Alan Karr, National Institute of Statistical Sciences (Modeling the Conductivity of Concrete); Hugh Barton and Woodrow Setzer, Environmental Protection Agency (Evaluating a Physiologically Based Pharmacokinetic Model); Tony Royal, Jenike and Johanson,

Inc. (Effect of Interstitial Gas on Powder Flow) and Pierre Maldague, Jet Propulsion Laboratory (Planning and Scheduling Strategy for Non-Deterministic Events) [for more details see CRSC-TR04-07 at http://www.ncsu.edu/crsc/reports/reports04.htm ]. Again SAMSI Graduate Fellows and Postdocs were heavily involved as local mentors for the participants. A number of the projects were excellent vehicles for illustration of the SAMSI focus on combining applied mathematical and statistical approaches to important scientific and engineering problems. A similar workshop (with new projects and presenters) is scheduled for July 26 - August 3 in 2004.

**Workshop on Mathematics Meets Biology: Epidemics, Data Fitting and Chaos:** In an effort to reach college teachers (typically at non Research I institutions) with the SAMSI message, SAMSI is joining with the Mathematical Association of America (MAA) in co-sponsoring one of their Professional Enhancement Programs (PREP, see http://www.maa.org/prep/) to see be held May 26-29, 2004 at the University of Louisiana at Lafayette. Participants will be 25-30 college teachers chosen from a national applicant pool. The SAMSI Associate Director for E&O along with several current SAMSI Graduate Fellows will be presenting one full day of the program; their material will include hands-on population modeling where combined deterministic and stochastic aspects of the models are needed to describe experimental data featuring both inter- and intra- individual characteristics.

**Mathematical and Experimental Modeling Course on NC-REN TV:** SAMSI sponsored an innovative course on closed circuit TV (the NC-REN TV network) to multiple campuses in North Carolina in the Fall, 2003 semester. This course (for a description see Appendix G) combined (in the SAMSI spirit) mathematical, statistical and experimental aspects of modeling of several physical and biological processes including thermal, mechanical and size-structured population systems. In addition to being given to a live audience on the NCSU campus, the course was live (and interactively) broadcast to classes at UNC-W, UNC-G, and ECSU. Students from the remote campuses came to the CRSC Lab at NCSU to carry out experiments with the NCSU students several times during the semester.

**Diversity:** See Section I.H for discussion of the efforts to achieve diversity.

**Courses:** See the program reviews in Section I.E and Appendix G for discussion of the SAMSI courses.

# 5 Planning and Hot Topics Workshops

## 5.1 For a program on High Dimensional Inference and Random Matrices

A meeting of potential main organizers was held at the American Institute of Mathematics, Palo Alto, February 29 and March 1. Many thanks to Helen Moore and Brian Conrey of AIM for all their help in hosting the meeting.

The participants were: Iain M. Johnstone, Statistics, Stanford University, Craig A. Tracy, Mathematics, University of California, Davis, Kenneth McLaughlin, Mathematics, University of North Carolina, Chapel Hill, Peter Bickel, Statistics, University of California, Berkeley, Douglas Nychka, National Center for Atmospheric Research, Hien T. Tran, Mathematics, North Carolina State University, James O. Berger, SAMSI, J. S. Marron, SAMSI

As a method of introduction, each participant gave an overview presentation of their related research and general interests. Discussion resulted in an agreement to focus on the main topics of: Extreme Sample Eigenvalues, Sample Eigenvectors, Empirical Distribution of Eigenvalues, Approximations of Empirical Data, Nonlinear and Topological Approaches to Dimension Reduction, Bayesian Version of Regularization in Large p Problems and Invariance Structure, Stochastic Evolution of Random Matrices, and Statistical Issues in EOFs in Climatology

A list of potential participants was also developed. There was general agreement that the program is viable, and that planning efforts should continue.

## 5.2 Hot Topics Workshop: Mathematical Sciences Research to Meet National Security Needs

The initial SAMSI hot topics workshop, on April 1-2, 2004, drew more than 30 attendees from industry, government and academia. The stage for working groups was set by four presentations:

- CDC Perspective: Lawrence Cox, National Center for Health Statistics
- DoD Perspective: Nancy Spruill, Office of the Secretary of Defense
- NSA Perspective: William Szewcyzk, National Security Agency
- Agroterrorism Perspective: Barrett Slenning, North Carolina State University

Two sets of working groups provided input to a potential SAMSI program on national defense and homeland security (NDHS). The first set matched the four application-oriented perspectives, while the second comprised four "proto" working groups for a SAMSI program: Anomaly Detection, Decisions, Modeling & Simulation, and Real Time Inference. There was unanimous support among the participants for a SAMSI program on NDHS in 2005-06. Initial discussions are underway with potential program leaders. Alan Karr is likely to be the directorate liaison and Sallie Keller-McNulty the NAC liaison.

A workshop report is being prepared.

# 6 Distinguished Lecture Series

This lecture series brought some of the worlds most prominent statistical and mathematical scientists to SAMSI. In addition to their very widely attended lectures, the distinguished visitors held highly useful discussions with SAMSI researchers. The list of distinguished lecturers for 2003-04, and the titles and abstracts of their talks, follow:

**Speaker:** *Margaret Wright,* Silver Professor and Chair of the Computer Science Department in the Courant Institute of Mathematical Sciences, New York University

**Date:** October 7, 2004

**Title:** Direct Search Methods: The Sound and the Fuss

**Abstract:** Direct search methods, which optimize without using derivatives, constitute a fascinating chapter, still being written, in the annals of optimization. Their history includes an initial heyday in the 1960s, a fall from grace in the 1970s and 80s, and a resurgence dating from the mid-1990s. Today's research on these methods has highlghted several interesting—in some cases, contentious—issues, ranging from the nature of convergence proofs and the associated assumptions to the proper role for direct search methods in modern optimization practice. We shall consider a selection of these issues, their current status, and their implications. And since this talk is being given at SAMSI, the speaker has high hopes for an improvisational segment devoted to connections between direct search methods and statistical sampling.

**Speaker:** *Jerome H. Friedman,* Professor of Statistics, Stanford University, and Leader of the Computation Research Group at the Stanford Linear Accelerator

**Date:** November 11, 2003

**Title:** Importance Sampling: An Alternative View of Ensemble Learning

**Abstract:** Learning a function of many arguments is viewed from the perspective of high-dimensional numerical integration. It is shown that many of the popular ensemble learning methods can be cast in this framework. In particular, bagging, boosting, and Bayesian model averaging are seen to correspond to Monte Carlo integration methods each based on different importance sampling strategies. This interpretation explains some of their properties and suggests modifications to them that can improve their accuracy and especially their computational performance.

**Speaker:** *Jonathan Chapman,* Professor of Mathematics and its Applications and Fellow of Mansfield College at Oxford University

**Date:** March 2, 2004

**Title:** A Hierarchy of Models for Type-II Superconductors

**Abstract:** One of the interesting aspects of superconductivity is the variety of models available to describe the phenomenon at different lengthscales, ranging from the microscopic theory of

Bardeen, Cooper & Schreiffer through the mesoscopic theories of London and Ginzburg & Landau, to the macroscopic Critical State theories such as the Bean model. The talk will explore the relationship between these different models by examining suitable asymptotic limits.

The basic building block in deriving this hierarchy is the superconducting vortex, which is a thin core of nonsuperconducting material circled by a superconducting electric current. Similar line singularities are found in other systems, for example, line vortices in an inviscid fluid, or Volterra dislocations in an elastic crystal. This raises the interesting question of whether analogous hierarchies of models exist in each of these other systems, and whether similar connections can be established between them.

**Speaker:** *Thomas G. Kurtz*, Paul Levy Professor of Mathematics and Statistics, University of Wisconsin - Madison

**Date:** April 6, 2004

**Title:** Particle Representations of Continuum Models

**Abstract:** Many stochastic and deterministic models are derived as the continuum limit of a discrete stochastic system as the size of the system tends to infinity. Discrete "particles" are each assigned a small mass and the limiting "mass distribution," typically characterized as a solution of a deterministic or stochastic partial differential equation, gives the desired model. A number of examples will be described in which keeping the discrete particles in the limit provides a useful tool for justifying the limit, analyzing the limiting model, and relating data to the limiting model. Possible examples include derivation of fluid models for internet protocols, many-server queueing approximations, models of stock prices set by infinitely many competing traders, consistency of numerical schemes for filtering equations, and models in population genetics; however, the speaker promises not to try to discuss all of these.

# F. Industrial and Governmental Participation

Government and industry participation in SAMSI programs and activities reflects broad interest in the SAMSI vision. The following summarizes participation during 2003-04.

*Data Mining and Machine Learning Program:* A variety of individuals participated in, or interacted significantly with, the program working groups. These included:

- More than 15 researchers and managers from General Motors took part in discussions associated with analysis of the vehicle sales data.
- The Bureau of Labor Statistics provided data and advice to support text mining to infer occupational categories for Census long-form answers.
- Scientists from ICAGEN (Research Triangle Park, NC), Metabolon (Research Triangle Park, NC) and the SAS Institute (Cary, NC) participated in the Bioinformatics Working Group throughout the year. The Metabolon relationship has generated one proposal to date; another is in progress. Others from Abbott Laboratories (Abbott Park, IL) and GlaxoSmithKline participated on a more limited basis.

*Internet:* The Scientific Committee includes 2 industrial members (AT&T Research and Avaya Labs). One of the theme problems of the Internet Tomography Workshop was a testbed developed by Avaya. The National Security Agency provided partial funding for the workshops on Internet Tomography and Sensor Networks. There many workshop participants from industry, including UAvaya Labs, CAIDA, Eurandom, IBM Research, Lucent Technologies Bell Labs, MCNC, NIST, Palo Alto Research Center, Sprint Labs, and the U.S. Navy,

*Multiscale Model Development and Control Design:*. The interdisciplinary nature of the Multiscale Program makes it highly amenable to industrial and government participation. During the Opening Workshop, there was significant participation, including invited presentations, by scientists from the Center for Disease Control (CDC), Lawrence Livermore National Lab, Los Alamos National Laboratory, NIEHS, and the US Army Research Office. The Workshop on Soft Matter Materials included participants from NIEHS and the US Army Research Office.

Additionally, research focused on model development and control design from combined stochastic/deterministic perspectives is being pursued with researchers from Sandia National Lab, Boeing, Kirtland AFB, and NASA Langley Research Center. This research will contribute directly to the Multiscale Program as well as the fundamental and technological research missions of the participating teams.

There was also strong industry/government/national laboratory participation through the NISS and NISS/SAMSI affiliates (see Section I.I.) and on the National Advisory Council, with members from Google, Microsoft and Los Alamos National Laboratory.

# G. Publications and Technical Reports

Because SAMSI programs have only been running for 6 months, we also indicate below technical reports that are under preparation.

I. INVERSE PROBLEM METHODOLOGY IN COMPLEX STOCHASTIC MODELS

**Publications and Technical Reports**

- Ackleh, A.S., K. Deng, C. Cole and H.T. Tran
  *"Existence-uniqueness and monotone approximation for an erythropoiesis age-structured model"*
  CRSC-TR03-21, April, 2003; submitted to J. Math. Anal. Appl.

- Ackleh, A.S., H.T. Banks, K. Deng, S. Hu
  *"Parameter estimation in a coupled system of nonlinear size-structured populations"*
  to be submitted.

- Adams, B.M., H.T. Banks, M. Davidian, H.D. Kwon, H.T. Tran, S.N. Wynne and E.S. Rosenberg
  *"HIV dynamics: Modeling, data analysis, and optimal treatment protocols"*
  CRSC-TR04-05, February, 2004; J. Comp. and Appl. Math., submitted.

- Adams, B.M., J.E. Banks, H.T. Banks and J.D. Stark
  *"Population dynamics models in plant-insect-herbivore-pesticide interactions"*
  CRSC-TR03-12, March, 2003, Revised, August, 2003; Math. Biosci., submitted.

- Banks, H.T., and J. Bardsley
  *"Well-posedness for systems arising in time domain electromagnetics in dielectrics"*
  CRSC-TR03-27, July, 2003; Communications in Applied Analysis, to appear.

- Banks, H.T., and J. Bardsley
  *"Parameter identification for a dispersive dielectric in 2D electromagnetics: Forward and inverse methodology with statistical considerations"*
  CRSC-TR03-48, December, 2003; Inverse Problems, submitted.

- Banks, H.T., D.M. Bortz, G.A. Pinter, and L.K. Potter
  *"Modeling and Imaging Techniques with Potential for Application in Bioterrorism,"*
  to appear as a Chapter in "Biomathematical Modeling Applications for Homeland Security" (H.T.Banks and C.C.Chavez, eds.), SIAM Frontiers in Applied Mathematics.

- Banks, H.T., N.G. Medhin and G.A. Pinter
  *"Multiscale considerations in modeling of nonlinear elastomers"*
  CRSC-TR03-42,October, 2003; J. Comp. Meth. Sci. and Engr., to appear.

- Banks, H.T., N.G. Medhin and G.A. Pinter
  *"Nonlinear reptation in molecular based hysteresis models for polymers"*
  CRSC-TR03-45, December, 2003; Quarterly Applied Math., to appear.

- Banks, H.T. and N.L. Gibson
  *"Well-posedness in Maxwell systems with distributions of polarization relaxation parameters"*
  CRSC-TR04-01, January, 2004; Applied Math. Letters, submitted.

- Banks, H.T. and G.A. Pinter
  *"A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue"*
  CRSC-TR04-03, January, 2004; SIAM J. Multiscale Modeling and Simulation, submitted.

- Banks, H.T., N.L. Gibson and W.P. Winfree
  *"Electromagnetic crack detection inverse problems using Terhertz interrogating signals"*
  CRSC-TR03-40, October, 2003; Inverse Problems, submitted.

- Banks, H.T., Y. Ma and L.K. Potter
  *"A simulation based comparison between parametric and semiparametric methods in a PBPK model"*
  J. Inverse and Ill-posed Problems, to be submitted.

- Bardsley, J.
  *"A bound-constrained Levenberg-Marquardt algorithm for a parameter identification problem in electromagnetics"*
  J. Inverse and Ill-posed Problems, submitted.

- Ma, Y. and M.G. Genton
  *"A Semiparametric Class of Generalized Skew-Elliptical Distributions"*
  Scandanavian Journal of Statistics to appear

- Ma, Y., M.G. Genton and M. Davidian
  *"Linear mixed effect models with semiparametric generalized skew-elliptical random effects"*
  to appear in *Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality,* Genton, M. G. (editor), Chapman & Hall / CRC, 2004.

- Ma, Y., M.G. Genton and A.A. Tsiatis
  *"Locally efficient semiparametric estimators for generalized skew-elliptical distributions."* J. American Statistical Association, submitted.

- Ma, Y. and A.A. Tsiatis
  *"Closed form semiparametric estimators for mixed models with complete sufficient statistic"*
  to be submitted.

- Tsiatis, A.A., and Y. Ma
  *"Locally efficient semiparametric estimators for functional measurement error models"*
  Biometrika, to appear.

II. CHALLENGES IN STOCHASTIC COMPUTATION

**Publications and Technical Reports**

- Carter, C., F. Wong and R. Kohn (2003)
  *"Efficient Estimation of Covariance Selection Models"*
  Biometrika, 90, 809-830.

- Chen, Y., I.H. Dinwoodie, A. Dobra and M. Huber (2003)
  *"Lattice Points, Sampling, and Contingency Tables"*
  Contemporary Mathematics (to appear)

- Chen, Y. and D. Small (2004)
  *"Testing the Rasch Model via Sequential Importance Sampling"*
  Psychometrika (to appear).

- Clyde, M. and E. George (2004)
  *"Model Uncertainty." Statistical Science (to appear)*

- Cripps, E., C. Carter and R. Kohn, R. (2004)
  *"Variable Selection and Covariance Selection in Multivariate Regression Models"*
  Handbook of Statistics: Bayesian Statistics: Modeling and Computation, (eds: C.R. Rao and D.K. Dey), Elsevier Press (to appear).

- Dinwoodie, I.H., L.F. Matusevich and E. Mosteig, E. (2004)
  *"Transform Methods for the Hypergeometric Distribution"*
  Statistics and Computing (to appear).

- Dinwoodie, I.H. (2003)
  *"Estimation of Parameters in a Network Reliability Model with Spatial Dependence." Submitted to Mathematics of Operations Research*

- Dinwoodie, I.H. and B. MacGibbon (2003)
  *"Exact Analysis of a Paired Sibling Study"*
  Submitted for publication

- Dobra, A., B. Jones, C. Hans, J.R. Nevins and M. West (2003)
  *"Sparse Graphical Models for Exploring Gene Expression Data"*
  Journal of Multivariate Analysis (to appear).

- Dobra, A., C. Tebaldi and M. West (2003)
  *"Bayesian Inference for Incomplete Multi-way Tables"*
  Submitted for publication.

- Huber, M., Y. Chen, A. Dobra, M. Nicholas and I.H. Dinwoodie (2003)
  "Monte Carlo Algorithms for Hardy-Weinberg Proportions"
  Submitted to Biometrics

- Ji, C. and B. Lee (2004)
  *"Central Limit Theorems in Computation of Option Prices with Stochastic Volatility Models"*
  Submitted to Mathematical Finance

- Jones, B., C. Carvalho, A. Dobra, C. Hans, C. Carter and M. West (2004)
  *"Experiments in Stochastic Computation for High-Dimensional Graphical Models"*
  Submitted to Statistical Science

- Lee, B. and C. Ji (2004)
  *"MCMC Calibration of Stochastic Volatility Models"*
  Submitted for publication.

- Molina, G., C.H. Han and J.P. Fouque (2003)
  *"MCMC Estimation of Multiscale Stochastic Volatility Models"*
  Submitted to Journal of Applied Econometrics

**Reports in Preparation**

- Cheng, A. and C. Ji (2004)
  "Gaussian Approximations in Option Pricing and Calibration of Stochastic Volatility Models"

- Clyde, M., G. Molina and M. Littman (2004)
  *"Bayesian Adaptive Stochastic Sampling for Variable Selection"*

- Hans, C., A. Dobra and M. West (2004)
  *"Shotgun Stochastic Search for Regression Model Uncertainty and Exploration"*

- Liang, F., R. Paulo, G. Molina, M. Clyde and J. Berger (2004)
  *"Gaussian Hyper-Geometric and Other Mixtures of g– Priors for Bayesian Variable Selection"*

- Paulo, R., G. Molina, C. Kohnen, M. Clyde and J. Berger (2004)
  *"Stochastic Computation in Bayesian Variable Selection"*

- Wong, F., C. Carter and R. Kohn (2004)
  *"Testing for Structure in the Inverse Covariance Matrix"*

III. LARGE-SCALE COMPUTER MODELS FOR ENVIRONMENTAL SYSTEMS

**Publications and Technical Reports**

- Anderson, D.M., R.M. McLaughlin, and C.T. Miller (2003)
  "*The Averaging of Gravity Currents in Porous Media"*
  In review: Physics of Fluids

- Caragea, P. (2003),
  *"Approximate likelihoods for spatial processes"*
  PhD dissertation, Department of Statistics, University of North Carolina-CH

- Culligan, K. A., D. Wildenschild, B. S. B. Christensen, W. G. Gray, and A. F. B. Tompson (2004)
  *"Interfacial Area Measurements for Unsaturated Flow Through a Porous Medium" to be submitted*

- Farthing, M. W., C. E. Kees, T. S. Coffey, C. T. Kelley and C. T. Miller (2003) "Efficient Steady-State Solution Techniques for Variably Saturated Groundwater Flow"
  Advances in Water Resources, 28 number 8, 833-849

- Fowler, K. R. and C. T. Kelley
  *"Pseudo-transient Continuation for Nonsmooth Nonlinear Equations"*
  Center for Research in Scientific Computation, CRSC-TR03-29, July 2003

- Fowler, K. R., and C. T. Kelley, C. E. Kees and C. T. Miller
  *"A Hydraulic Capture Application for Optimal Remeidation Design"*
  Center for Research in Scientific Computation, CRSC-TR04-04
  To appear in the proceedings of Computational Methods in Water Resources XIV, February, 2004

- Fuentes, M. (2003)
  *"Testing for separability of spatial-temporal covariance functions"*
  Submitted for publication

- Gray, W. G. and C. T. Miller (2004)
  *"An Examination of Darcy's Law for Flow in Variable Porosity Porous Media" submitted to Environmental Science & Technology*

- Gray, W. G. and C. T. Miller (2004)
  *"Thermodynamically Constrained Averaging Theory Approach to Modeling of Flow in Porous Media: 1. Motivation and Overview"*
  submitted to Advances in Water Resources

- Kasibhatla, P., A. Arellano, J. A. Logan, P. I. Palmer and P. Novelli (2002)
  "Top-down estimate of a large source of atmospheric carbon monoxide associated with fuel combustion in Asia"
  Geophys. Res. Lett., 29 (19), 1900

- Kavanagh, K.R., C.T. Kelley, C.T. Miller, M.S.C. Reed, C.E. Kees, and R.W. Darwin (2003) *"Solution of a Well-Field Problem with Implicit Filtering"*
  In review: Optimization and Engineering

- Kavanagh, K. R., C. T. Kelley, C. T. Miller, C. E. Kees, R. W. Darwin, J. P. Reese, M. W. Farthing and M. S. C. Reed (2004)
  *"Solution of a Well-Field Design Problem with Implicit Filtering"*
  Optimization and Engineering (to appear)

- Kelley, C. T., K. R. Fowler and C. E. Kees
  *"Simulation of Nondifferentiable Models for Groundwater Flow and Transport"*
  Center for Research in Scientific Computation, CRSC-TR03-47, December 2003. To appear in the proceedings of Computational Methods in Water Resources XIV.

- Le, N. D., L. Sun and J. V. Zidek
  "Designing Networks for Monitoring Multivariate Environmental Fields Using Data With Monotone Pattern"
  SAMSI, 2003-5, May 23, 2003

- Mayer, A. S., C. T. Kelley and C. T. Miller (2004)
  *"Optimal design for problems involving flow and transport in saturated subsurface systems" Advances in Water Resources, 25, 8-12*

- McLaughlin, Richard, David Adalsteinnson, Nicole Abaid, and Akua Aguapong
  *"An Internal Splash: Levitation of Falling Sphere's in Stratified Fluids"*
  Submitted for publication

- Pan, C., M. Hilpert and C. T. Miller (2004)
  *"Lattice-Boltzmann Simulation of Two-Phase Flow in Porous Media"*
  Water Resources Research, 40 (1)

- Pan, C., J. F. Prins and C. T. Miller (2004)
  *"A High-Performance Lattice Boltzmann Implementation to Model Flow in Porous Media" Computer Physics Communication, 158 (2), 89-105*

- Serre, M.L., G. Christakos, H. Li, and C.T. Miller (2003)
  *"A BME Solution of the Inverse Problem for Saturated Groundwater Flow"*
  In press: Stochastic Environmental Research and Risk Assessment

- Tebaldi, C., R.L. Smith, D. Nychka and L.O. Mearns
  *"Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles"*
  National Center for Atmospheric Research, January 2004

- Zidek, J.V., J. Meloche, G. Shaddick, C. Chatfield and R. White
  *"A Computational Model for Estimating Personal Exposure to Air Pollutants with Application to London's PM10 in 1997"*
  SAMSI, 2003-3, March 27, 2003

**Reports in Preparation** (as of April, 2003)

- Camassa, Roberto, Ken McLaughlin, Rich McLauglin, and James Bonn
  *"Effective mixing coefficients for passive transport"*

- Chang, H, Fu, H, Le, ND, Zidek, J.V.
  *"Perspectives on Designing Environmental Monitoring Networks for Measuring Extremes"*

- Fu, A, Le, N.D., Zidek, J.V.
  *"A Statistical Characterization of a Simulated Maximum Annual Rainfall Field Over Canada"*

- Kavanagh, K. R., C. T. Kelley, C. T. Miller, M. Reed, C. Kees and R. Darwin
  *"A Comparision of Optimization Methods for Problems Involving Flow and Transport Phenomena in Saturated Subsurface Systems"*

- Kavanagh, K. R. and C. T. Kelley
  *"Temporal Error Control and Pseudo-transient Continuation for Nonsmooth Dynamics"*

- McLaughlin, Rich, Dan Anderson, and Casey Miller
  *"Gravity currents in heterogeneous porous media systems"*

- Rich McLaughlin, Anne Bourlioux, and Roberto Camassa
  *"Non-Gaussian PDFs in time varying shear layers"*

- Natvig, Bent and Tvete, Ingunn Fride (2003)
  *"Bayesian hierarchical modelling of spatial and temporal dependencies between earthquakes"*

- Zheng, Xiaoyu and M. Gregory Forest
  *"On the strength of monodomain attractors for sheared nematic polymers"*

IV. DATA MINING AND MACHINE LEARNING

**Publications and Technical Reports**

- Banks, D., House, L., Arabie, P., McMorris, F. R., and Gaul, W., eds. (2004).
  *"Classification, Clustering, and Data Mining"*
  Springer-Verlag, Heidelberg.

- Beecher, C., Cutler, A., House, L., Lin, X., Truong, Y., and Young, S. (2004).
  *"Learning a complex metabolomic dataset using random forests and support vector machines."*
  Submitted to SIGKDD 2004.

- Fokoué, E. (2004).
  *"Sparsity through prevalence estimation."*
  To be submitted to the J. Machine Learning Res.

- Fokoué, E. (2004).
  *"Sparsity through prevalence estimation."*
  To be submitted to the \textit{J. Machine Learning Res.

- Hawkins, D. M., Wolfinger, R. D., Liu, L., and Young, S. S. (2003).
  *"Exploring blood spectra for signs of ovarian cancer."*
  Chance, 16, 19-23.

- House, L., and Banks, D. (2004).
  *"Cherry picking as a robustness tool."*
  In Classification, Clustering, and Data Mining, Springer-Verlag, Heidelberg, pp. 197-208.

- House, L., and Banks, D. (2004).
  *"Robust multidimensional scaling."*
  Submitted to Proceedings of COMPSTAT'04.

- Lin, X., and Yu, Z. (2004).
  *"Degenerate expectation-maximization algorithm for local dimension reduction."*
  In Classification, Clustering, and Data Mining,
  Springer-Verlag, Heidelberg, pp. 259-268.

- Liu, L., Hawkins, D. M., Ghose, S., and Young, S. S. (2004)
  *"Robust singular value decomposition analysis of microarray data"*
  Proc. Nat. Acad. Sci. (100) 13167-13172

- Simmons, S., Lin, X., Beecher, C., Truong, Y., and Young, S. (2004).
  *"Active and passive learning to explore a complex metabolomics dataset."*
  In Classification, Clustering, and Data Mining,
  Springer-Verlag, Heidelberg, pp. 447-456.

- Young, S. S., and Ge, N., (2004).
  *"Design of diversity and focused combinatorial libraries in drug discovery."*
  To appear in Current Opinions in Drug Design and Development

- Young, S. S., Wang, M., and Gu, F. (2003).
  *"Design of diverse and focused combinatorial libraries using an alternating algorithm."*
  J. Chem. Info. Comp. Sci. (43) 1916-1921.

**Reports in Preparation**

- Fokoué, E., Sun, D. and Goel, P. (2004).
  *"A new hierarchical prior structure for the relevance vector machine."*

- Young, S. S., Feng, J., and Sanil, A.
  *"PharmID: Pharmacophore identification using Gibbs sampling."*

- Zhang, H., Jeongyoun Zhang, J., Lin, X., and Park, C. (2004).
  *"Support vector machine with feature selection using a nonconcave penalty."*

V. NETWORK MODELING FOR THE INTERNET

**Publications and Technical Reports**

- Park, C., Godtliebsen, F., Taqqu, M., Stoev, S. and Marron, J. S.
  *"Visualization and Inference Based on Wavelet Coefficients, SiZer and SiNos"*
  SAMSI Technical Report No. 2004-10.

- Park, C., Hernandez-Campos, F., Marron, J. S., Rolls, D. and Smith, F. D.
  *"Long-Range-Dependence in a Changing Internet Traffic Mix"*
  SAMSI Technical Report No. 2004-9.

- Stoev, S., Taqqu, M., Park, C. and Marron, J. S.
  *"LASS: a Tool for the Local Analysis of Self-Similarity"*
  SAMSI Technical Report No. 2004-7.

- Stoev, S., Taqqu, M., Park, C. and Marron, J. S.
  *"Strengths and Limitations of the Wavelet Spectrum Method in the Analysis of Internet Traffic"*
  SAMSI Technical Report No. 2004-8.

- Xu, P., Devetsikiotis, M. and Michailidis, G.
  *"Adaptive Scheduling using Online Measurements for Efficient Delivery of Quality of Service"*
  SAMSI Technical Report No. 2004-12.

- Xu, P., Devetsikiotis, M. and Michailidis, G.
  *"Online Scheduling for Resource Allocation of Differentiated Services: Optimal Settings and Sensitivity Analysis"*
  SAMSI Technical Report No. 2004-13.

- Zhu, Z. and Taqqu, M. S.
  *"Impact of the sampling rate on the estimation of the parameters of fractional Brownian motion"*
  SAMSI Technical Report No. 2004-14.

**Reports in Preparation**

- Abry, P. and Pipiras, V.
  *"Wavelet-based synthesis of the Rosenblatt process"*

- Dinwoodie, I. H., Mosteig, E., Gamunid, E.
  "Algebraic Equations for Blocking Probabilities in Asymmetric Networks"

- Hernandez-Campos, F., Le, L., Marron, J. S., Park, C., Park, J., Pipiras, V. Smith, F. D., Smith, R. L., Trovero, M. and Zhu, Z.
  *"Long Range Dependence Analysis of Internet Traffic"*

- Park, C., Hernandez-Campos, F., Marron, J. S. and Jeffay, K.
  *"Thresholded Log-Log Correlation Analyses of TCP Characteristics"*

- Park, C., Marron, J. S. and Rondonotti, V.
  *"Dependent SiZer: Goodness of Fit Tests for Time Series Models"*

- Park, C., Veitch, D., Shen, H., Hernandez-Campos, F., and Marron J. S.
  *"Semi experiment analysis of the shifting knee wavelet spectrum"*

- Park, J. and Park C.
  *"Robust H estimation, automatic choice of parameters"*

- Pipiras, V.
  *"On the use and usefulness of wavelet-based simulation of fractional Brownian motion"*

- Piparas, V. and Taqqu, M. S.
  *"Identification of periodic and cyclic fractional stable motions"*

- Piparas, V. and Taqqu, M. S.
  *"Integral representations of periodic and cyclic fractional stable motions"*

- Piparas, V. and Taqqu, M. S.
  *"Semi-additive functionals and cocycles in the context of self-similarity"*

- Rolls, D., Michailidis, G. and Hernandez Campos. F.
  *"Queueing Analysis of Network Traffic"*

- Smith, R. L., Taqqu, M., Shen, H., Park, J., Zhu, Z. and Park, C.
  *"Change Points and Long Range Dependence"*

- Zhu, Z., Shen, H., Park, C., Hernandez-Campos, F and Marron, J. S.
  *"Shot noise model, start times, micro-bursts"*

VI. MULTISCALE MODEL DEVELOPMENT AND CONTROL DESIGN

**Publication and Technical Reports**

- Ball, B.L., and R.C. Smith
  *"A Stress-Dependent Hysteresis Model for PZT-Based Transducers"*
  SAMSI Tech Report 2004-6; Proceedings of the SPIE, Smart Structures and Materials 2004, to appear.

- Edmonds,.Jr., B., J. Ernstberger, K. Ghosh, J. Malaugh, D. Nfodjo, W. Samyono, X. Xu, D. Dausch, S. Goodwin and R.C. Smith
  *"Electrostatic Operation and Curvature Modeling for a MEMS Flexible Film Actuator"*
  SAMSI Tech Report 2004-4;Proceedings of the SPIE, Smart Structures and Materials 2004, to appear

- Hatch, A.G., R.C. Smith and T. De
  *"Model Development and Control Design for High Speed Atomic Force Microscopy"*
  SAMSI Tech Report 2004-3; Proceedings of the SPIE, Smart Structures and Materials 2004, to appear.

- Raye, J.K., and R.C. Smith
  *"A Temperature-Dependent Hysteresis Model for Relaxor Ferroelectric Compounds"*
  SAMSI Tech Report 2004-5; Proceedings of the SPIE, Smart Structures and Materials 2004, to appear.

# H. Efforts to Achieve Diversity

From the beginning, SAMSI has focused on achieving diversity. This begins with composition of advisory committees. On the National Advisory Committee, 3 of 11 are women, including the Co-Chair, Margaret Wright, and one member is Hispanic. On the Local Advisory Committee, 1of 8 is a woman and two are African-American. On the Education and Outreach Committee, there were two African-Americans and one woman on the 7 member committee.

Specific efforts, and successes, of each of the programs towards achieving diversity are indicated below.

***Data Mining and Machine Learning:*** Women and new researchers were well-represented throughout the program, with the possible exception of the invited speakers at the Kickoff Workshop, only one of whom was female. However, four of the seven talks at the new researchers session were by women. Approximately one-half of visitors were women, and speakers at the January--February mid-term workshops included a number of new researchers (among them, graduate students and postdoctorals).

Despite vigorous efforts, participation by under-represented minorities did not attain SAMSI goals. One of the two DMML postsdoctoral fellows is black, as is one Faculty Fellow. Note, however, that minority participation at the two undergraduate 2-day workshops, which were based on DMML material, was very strong, as detailed below.

***Network Modeling for the Internet:*** A number of women have been major players at all phases of this program, as noted below. Furthermore, strenuous efforts have been taken to address diversity, as detailed below.

network related research; Jeoungyoun Ahn, a PhD student at UNC; and Rima Izem, a PhD student at UNC.

***Multiscale Model Development and Control Design:*** Two significant goals of the program are to make it as widely accessible to students and new researchers as possible and to recruit a diverse range of participants. Both goals are being addressed through aggressive solicitation by the program leaders and committee via personal and research contacts as well as formal symposia and presentations. For example, the majority of participants who attended the Opening Workshop were notified by either the organizers or committee. Of the 82 attendees on the second day of the workshop, 19 were women, 4 were African American and 1 was Hispanic. Similar demographics were observed during the remainder of the workshop as well as at the second workshop.

***Education and Outreach Program:*** SAMSI continues to use its E&O program to enhance its diversity efforts by active recruitment of under represented participants. Special efforts are made to recruit from HBCU's for all programs. During the past year, participation breakdowns include:

- Undergraduate Workshop (June, 2003): Out of 15 participants, 11 were female and 4 were African American.
- Graduate Workshop (July, 2003): Out of 37 participants, 11 were female and 2 were African American.
- Undergraduate Workshop (November, 2003): Out of 30 participants, 20 were female and 5 were African American.
- Undergraduate Workshop (February, 2004): Out of 21 participants, 12 were female, 2 were African American.

# I. External Support and Affiliates

## 1. Additional Funding

**Kenan Foundation:** provided $50,000 of supplementary support.

**Internet Program:** The workshops on Internet Tomography and on Sensor Networks were jointly sponsored by SAMSI and the National Security Agency. The National Security Agency provided $15,000 for this purpose. J. S. Marron, SAMSI, was P.I. on the grant and Deborah Estrin, University of California, Los Angeles and Robert Nowak, University of Wisconsin, were co-P.I.s.

**Data Mining and Machine Learning Program:** in-kind support from industry and government was provided as follows:

- More than 15 researchers and managers from General Motors took part in discussions associated with analysis of the vehicle sales data.
- The Bureau of Labor Statistics provided data and advice to support text mining to infer occupational categories for Census long-form answers.
- Scientists from ICAGEN (Research Triangle Park, NC), Metabolon (Research Triangle Park, NC) and the SAS Institute (Cary, NC) participated in the Bioinformatics Working Group throughout the year. The Metabolon relationship has generated one proposal to date; another is in progress. Others from Abbott Laboratories (Abbott Park, IL) and GlaxoSmithKline participated on a more limited basis.

## 2. Affiliates Program

**Background:** The NISS Affiliates Program and NISS/SAMSI University Affiliates Program constitute the largest such programs among the DMS-funded mathematical sciences research institutes. Two Assistant Directors of NISS have major responsibility for operation of these programs.

As a benefit of membership, NISS Affiliates and NISS/SAMSI University Affiliates may receive reimbursement for expenses to attend SAMSI workshops as well as NISS events. Through meetings and other activities, the NISS Affiliates and NISS/SAMSI University Affiliates inform the development of SAMSI programs. To illustrate, the DMML Program for 2003–04 exists to a significant degree because of interest expressed at an Affiliates Planning Meeting in March of 2002.

NISS Affiliates and NISS/SAMSI Affiliates are listed below:

- **Corporations:** Amgen, Thousand Oaks; Avaya Labs, Basking Ridge, NJ; General Motors, Detroit, MI; GlaxoSmithKline, Research Triangle Park, NC and Collegeville, PA; ICAGEN, Research Triangle Park, NC; Merck & Company, West Point, PA; MetaMetrics, Durham, NC; Pfizer Inc., Groton,

CT; RTI International, Research Triangle Park, NC; SAS Institute, Cary, NC; SPSS, Chicago, IL; Telcordia Technologies, Morristown, NJ

- **Government Agencies and National Laboratories:** Bureau of Labor Statistics, Washington, DC; US Census Bureau, Washington, DC; Los Alamos National Laboratory National Agricultural Statistics Service, Fairfax, VA; National Center for Education Statistics, Washington, DC; National Center for Health Statistics, Hyattsville, MD; National Institute of Standards and Technology, Gaithersburg, MD; National Security Agency, Ft. George G. Meade, MD; Pacific Northwest National Laboratory, Richland, WA

- **NISS/SAMSI University Affiliates:** UCLA, Department of Statistics and Statistical Consulting Center; Carnegie Mellon University, Department of Statistics; Duke University, Institute of Statistics and Decision Sciences and Department of Mathematics; Emory University, Department of Biostatistics; Florida State University, Department of Statistics; University of Georgia, Department of Statistics; University of Illinois Urbana-Champaign, Department of Statistics; University of Iowa, Department of Statistics; Iowa State University, Department of Statistics Johns; Hopkins University, Department of Statistics and Applied Mathematics; University of Maryland Baltimore County, Department of Mathematics and Statistics; University of Michigan, Departments of Statistics and Biostatistics; University of Missouri-Columbia, Department of Statistics; North Carolina State University, Department of Mathematics; North Carolina State University, Department of Statistics; University of North Carolina at Chapel Hill, Department of Biostatistics; University of North Carolina at Chapel Hill, Department of Mathematics; University of North Carolina at Chapel Hill, Department of Statistics and Operations Research; Oakland University, Department of Mathematics and Statistics; Ohio State University, Department of Statistics; Pennsylvania State University, Department of Statistics; Rice University, Department of Statistics; Rutgers University, Department of Statistics; Southern Methodist University, Statistical Science Department; Stanford University, Department of Statistics; Texas A&M University, Department of Statistics

**Affiliate Participation:** Every SAMSI program and event during 2003–04 had strong Affiliate participation, nearing one-half of attendees at some workshops. Expenditures from Affiliates Reimbursement Account expenditures to attend SAMSI events exceeded $20,000.

Participation in the DMML and Internet Traffic programs by corporate, government and national laboratory affiliates was especially deep. Examples include:

- *General Motors (GM)* provided testbed databases and ongoing advice and assistance from more than a dozen researchers and managers for the DMML program.

- Personnel from *Avaya Labs* wwere major participants in the Internet Tomography thrust of the Internet Traffic program.

- The *Bureau of Labor Statistics (BLS)* provided data and advice to the DMML program to support text mining to infer BLS occupational categories for Census long-form answers.

- Researchers from *SAS Institute* participated throughout the year in the Bioinformatics working group of the DMML program.

- The April 1–2 Hot Topics Workshop on Mathematical Sciences Research to Meet National Security Neeeds will include participants from the *NCHS*, *NSA*, *Telcordia Technologies* and several university affiliates.

**Plans for the Future:** At the annual affiliates planning meeting, held at NISS/SAMSI on March 5, 2004, affiliates were invited to submit ideas or proposals for future SAMSI programs, and briefed on programs planned for 2004–05 and 2005–06. Although there was strong interest in all programs, support for the social sciences program in 2004–05 was especially widespread among corporate, government and national laboratory affiliates. In particular, one affiliate has volunteered to organize, and several others to assist, a workshop or summer school on social networks as part of the program.

# J. Advisory Committees

The four advisory/oversight committees of SAMSI are as follows:

- The Governing Board (GB), which oversees SAMSI's administration, finances, evaluation and partner organization relationships. The GB meets with the Directorate twice a year. The SAMSI Director also has a conference call with the GB Chair and/or GB every other week.

- The National Advisory Committee (NAC) consists of leading national scholars, and is the primary external input into program choice and development. The NAC met with the Directorate, at SAMSI, on October 25, 2003, to review the progress in the current programs and to consider the pre-proposals and proposals that had been submitted for programs in future years. In addition, there are frequent e-mails to the NAC asking for advice concerning developing or new programs. Finally, a member of the NAC serves as a Liaison with each of the Scientific Committees of the major SAMSI programs.

- The Local Development Committee (LDC) consists of leading local scholars, and has a crucial role to play in the involvement of local individuals in SAMSI programs, including the Faculty Release Fellows, the Graduate Associates, and the University Fellows. The LDC has met with the Directorate on October 27, 2003.

- The Chairs Committee, which consists of the chairs of the following departments at the partner universities:
    - Duke: Biostatistics and Bioinformatics, Institute of Statistics and Decision Sciences, Mathematics
    - NCSU: Mathematics, Statistics
    - UNC: Biostatistics, Mathematics, Statistics and Operations Research

  Note that the Chairs are also ex officio members of the LDC. Meetings with the Chairs were held before (and during) the LDC meeting mentioned above.

The membership of each of these committees during the past year is given in the table on the following page.

| Committee | Name | Affiliation | Field |
|---|---|---|---|
| Governing Board | John Harer | Duke, As. Provost | Mathematics |
| | Douglas Kelly | UNC, Dean | Statistics |
| | Jon Kettenring | Telcordia | Math Sciences |
| | Daniel Solomon | NCSU, Dean | Statistics |
| National Advisory Committee | Mary Ellen Bock | Purdue | Statistics |
| | Peter Bickel (Co-Chair) | UC Berkeley | Statistics |
| | Lawrence Brown | Pennsylvania | Statistics |
| | Carlos Castillo-Chavez | Arizona State | Math & Stat |
| | David Heckerman | Microsoft | CS & Statistics |
| | John Lehoczky | Carnegie Mellon | Probability |
| | Sallie Keller-McNulty | LANL | Statistics |
| | Daryl Pregibon | Google, Inc | Computer Science |
| | G.W. Stewart | Maryland | Computer Science |
| | Philippe Tondeur | Illinois | Mathematics |
| | Margaret Wright (Co-Chair) | NYU | CS |
| Local Development Committee | Gregory Forest | UNC | Mathematics |
| | Jean-Pierre Fouque | NCSU | Mathematics |
| | Jacqueline Hughes-Oliver | NCSU | Statistics |
| | Thomas Kepler | Duke | Bioinformatics |
| | Andrew Nobel | UNC | Statistics |
| | John Trangenstein | Duke | Mathematics |
| | David Banks | Duke | Bioinformatics |
| | Lloyd Edwards | UNC | Biostatistics |
| Chairs Committee | Vidyadhar Kulkarni | UNC | Operations Research |
| | Clarence E. Davis | UNC | Biostatistics |
| | Bernard Mair | NCSU | Mathematics |
| | David Morrison | Duke | Mathematics |
| | Sastry Pantula | NCSU | Statistics |
| | Dalene Stangl | Duke | Statistics |
| | William Wilkinson | Duke | Biostatistics |
| | Christopher Jones | UNC | Mathematics |

# II. Special Report: Program Plan

## A. Programs for 2004-2005

I. GENOMES TO GLOBAL HEALTH: THE COMPUTATIONAL BIOLOGY OF INFECTIOUS DISEASE (full year)

The eradication of naturally-occurring smallpox in 1980 marked the culmination of two centuries of work and stands as one of humankind's greatest achievements. This singular event, and the steady progress made leading up to it, contributed to an optimism that infectious disease would no longer be a threat to the industrialized world. The next year, a strange new illness appeared that would later become known as AIDS. Infectious disease remains a major cause of suffering and death among people in the developing world; globalization ensures that the developed world is only marginally safer. The emergence of HIV and SARS as novel viral agents, readily transported from rural and underdeveloped areas to the west, resulting in human and economic devastation in both regions underscores the seriousness of this concern. Some African countries such as Swaziland have adult HIV infection rates approaching 40%. The World Health Organization estimates economic cost of the SARS epidemic at about $30 billion. The WHO further estimates that 3000 African children die each day from malaria. Drugs once effective against the infectious agent, *Plasmodium falciparum* no longer work. Similarly, tuberculosis, once a fearsome killer of children in the developed world, and then virtually eradicated, has made a stunning recovery; drug-resistant *Mycobacterium tuberculosis* is now a significant health concern in New York City as well as in Bombay.

The near-completion of the human genome project brings with it the promise of a more complete understanding of human disease, including infectious disease, but the path from the genome sequences of the human hosts and their microbial pathogens is exceedingly complex and will require significant contributions from the mathematical sciences for its elucidation.

The primary aims of this year of research and study are to identify those areas where mathematical innovation may have the greatest impact on the basic science and medicine of infectious disease, to progress materially toward major research efforts in these areas, to establish a greater sense of community among the researchers with skills and interests in these areas, and to contribute to the training of the next generation of mathematically literate biomedical researchers, originating in both the biological and the mathematical sciences.

We envision the program as consisting of three intertwining strands representing different scales of resolution: genomics & molecular biology, cellular dynamics & physiology, and epidemiology & global health. Each of these strands represents a specific window onto the world; a major concern of ours will be to integrate the views revealed from these disparate perspectives, to explore novel multiscale approaches to the fundamental problems.

**Program Activities:**

*Foundational Workshop:* We will kick off the year's activities with the Foundational Workshop, 18-22 September 2004. Tutorials covering basic material in immunology, microbiology, statistics and mathematical modeling will begin at noon on Saturday the 18th and continue through lunch on Sunday. The research portion of the workshop begins after lunch Sunday and ends at noon on Wednesday. We will feature talks by prominent researchers representing the three major strands mentioned above, and the major action item for the workshop participants is to define a set of topics around which to structure the ongoing activities of the working groups.

*Working groups:* Working groups will meet throughout the year, and will consist of local faculty members, postdocs, grad students and resident visiting scholars as well as short-term visitors. The focus topics of these working groups will be developed in the foundational workshop and refined over the course of each semester. Examples of such topics include influenza evolution, molecular evolution of mycobacterial pathogenicity, signaling in leukocytes, ecology of commensalisms and pathogenicity, vaccine design. individual-based models for epidemiology, global patterns of disease spread, etc.

*Transitional workshop:* The formal program will draw to a close with a workshop in the late spring/early summer of 2005. The primary aims of this workshop will be to disseminate the results of the working groups over the course of the year and to strategize about preserving the momentum of built during the year, and preserving the community built in that time.

**Graduate courses**. In addition to the course in mathematical models of signal transduction, we are developing and will offer a full-semester course in *Computational Immunology and Immunogenomics* with Tom Kepler and Lindsay Cowell in the fall semester and *Mathematical Epidemiology* with Alun Lloyd. The signal transduction course will be half-semester, with the other half devoted to *Microbial Genomics*.

**Program Leadership:**

*Principal Organizer:* Thomas B. Kepler, Departments of Biostatistics & Bioinformatics and Immunology, Duke University

*Co-organizer:* Denise Kirschner, Department of Microbiology and Immunology. University of Michigan

*Co-organizer:* Lindsay Cowell, Department of Biostatistics & Bioinformatics, Duke University

*Local Scientific Committee*

| Name | Discipline | Affiliation |
|---|---|---|
| Atchley, William | Genetics | NCSU |
| Elston, Tim | Applied Mathematics | UNC |
| Nobel, Andrew | Statistics | UNC |
| Schmidler, Scott | Statistics and Decision Sciences | Duke |
| Thorne, Jeff | Statistical Genetics | NCSU |

*Global Scientific Committee*

| Name | Discipline | Affiliation |
|---|---|---|
| Anderson, Roy | Theoretical Epidemiology | Imperial College, London |
| Antia, Rustom | Biology | Emory |
| Bergstrom, Carl | Biology | University of Washington |
| Casadevall, Arturo | Microbiology and Immunology | Albert Einstein College of Medicine |
| Castillo-Chavez, Carlos | Biological Statistics and Computational Biology | Cornell |
| Gupta, Sunetra | Mathematical Epidemiology | Oxford |
| Perelson, Alan | Theoretical Biology and Biophysics | Los Alamos |
| Tan, Man-Wah | Immunology | Stanford |
| Wong, Wing | Statistics | Harvard |

## Participants:

*Visiting Scholars:* We are extremely fortunate to have Byron Goldstein, of Los Alamos National Laboratories joining us as the SAMSI University Fellow for this program. His expertise is in mathematical modeling of signal transduction in cells of the immune system, and he will be teaching a course on that topic during his residency in the Spring semester. Other scholars with expertise in diverse topics in immunology, microbiology, epidemiology, mathematics, statistics and computational biology, will be joining us for periods of varying duration. Most will be invited, but we are soliciting applications as well.

*Postdoctoral fellows:* SAMSI will appoint two postdoctoral fellows for this program, both of whom will continue with research in the second year of their appointment in the Duke University Laboratory of Computational Immunology.

*Faculty Release Fellows* from the partner universities are Tom Kepler, Lindsay Cowell, Tim Elston, Andrew Nobel, Scott Schmidler, all mentioned above, and Alun Lloyd (Mathematics, NCSU).

*Graduate Fellows* so far selected are Ben Cooke (Mathematics, Duke), Laura Ellwein (Mathematics, NCSU), Abel Rodriguez (Statistics, Duke), and Xiao Want (Statistics, UNC).

### II. LATENT/HIDDEN VARIABLE MODELS IN THE SOCIAL SCIENCES (full year)

Latent and hidden variables (LHV) are ubiquitous in the social sciences. In settings as diverse as intelligence or socioeconomic status, many variables cannot be directly measured. Factor analysis, latent class analysis, structural equation models, error-in-variable models, and item response theory illustrate models that incorporate latent variables. The SAMSI LHV program will take a broad look at latent variables and measurement error.

LHV program activities will include research, workshops and interaction with SAMSI Education/Outreach programs.

The research component of the program will operate similarly to the 2002--03 Stochastic Computation and 2003--04 Data Mining and Machine Learning programs, with several working groups focusing on specific issues of theory and methodology, many of which will be framed by particular testbed data sets. The precise foci of the Working Groups will be identified in conjunction with the September, 2004 Opening Workshop. Based on discussions of the scientific committee and with program participants, current candidates for Working Group foci are:

- Potential of causality
- Multilevel models
- Longitudinal data
- Relationships between hierarchical models and structural equation models
- Categorical variables in LHV models.

**Workshops:** Two major workshops are planned:

1. *Tutorials and Opening Workshop*, September 11-15, 2004. Following lessons learned from 2003-04 workshops, the format will be highly participatory, with Birds-of-a-Feather Sessions, a poster session and associated session of two-minute Poster Sales Talks, a Second Chance Seminar at which anyone can talk, a New Researchers Session, and Working Group Meetings at SAMSI.

2. *Closing Workshop*, May 19--21, 2005, at which findings of the program will be presented to the community, and follow-on activities catalyzed.

It is also likely that there will be several Mid-Program Focused Workshops, on topics and at dates to be announced. In addition, the National Program for Complex Data Structures in Canada has a social sciences initiative and is working with SAMSI to create a joint workshop for newer researchers in the area.

**Courses:** A SAMSI course on Longitudinal Modeling will be taught by Lloyd Edwards as part of the program.

**Program Leaders:** Kenneth A. Bollen (Sociology, University of North Carolina at Chapel Hill; Chair), James J. Heckman (Economics; University of Chicago), Alan F. Karr (NISS and SAMSI), and Susan A. Murphy (Statistics; University of Michigan).

**Participants:** A SAMSI-University Fellow for the program is currently being recruited. A number of people are planning lengthy research visits, including Maria Jesus Bayarri (Statistics, Valencia), Michael Browne (Psychology, Ohio State University) and Anders Skrondal (Sociology, Norway). A number of local participants will be extensively involved in the program, including Jerry Reiter (Statistics, Duke) and David Banks (Statistics, Duke). We expect that many additional participants, international, national and local, will come forward and be incorporated into the program on an ongoing basis.

At the March 5, 2004 Annual Planning Meeting of the NISS Affiliates and NISS/SAMSI University Affiliates, support for the social sciences program in 2004-05 was widespread,

especially among corporate, government and national laboratory affiliates. In particular, one affiliate has volunteered to organize, and several others to assist, a workshop or summer school on social networks as part of the program.

*Postdoctoral Fellows* for the program are

*Faculty Release Fellows* from the partner universities are Kenneth A. Bollen (Sociology, UNC), Lloyd Edwards (Biostatistics, UNC), Subhashis Ghosal (Statistics, NCSU), and Negash Medhin (Mathematics, NCSU).

*Graduate Fellows* so far selected are John Hipp (Sociology, UNC) and Satkartar Kinney (Statistics, Duke).

## III. Mathematical Challenges and New Directions for Data Assimilation in Geophysical Systems (Spring, 2005)

The issue of assimilating data into models arises in all scientific areas that enjoy a profusion of data. In its broadest sense, it is the subject that arises at the meeting point of data and models. Technology has driven the advances on both sides of the equation: new techniques of measurement have led to an enormous surge in the amount of available data and ever faster computers have given us the capability of new levels of computational modeling. The development of effective methods of data assimilation must now be viewed as one of the fundamental challenges in scientific prediction. Nevertheless, the part of the scientific community interested in these issues has been limited. With the growing need for good methods and the advances in computational and observational capabilities, we now have a tremendous opportunity to bring the relevant scientific areas together in a focused effort aimed at developing new approaches, understanding the underlying issues and testing implementations of new schemes.

The problem of assimilating data into a geophysical system is both fundamental in that it aims at the estimation of an unknown, true state and challenging as it does not naturally afford a clean solution. We shall focus on state estimation for systems related to the atmosphere and oceans. It has two equally important elements: observations and computational models. Observations measured by instruments provide direct information of the true state, whether they are taken in situ or by remote sensing. Such observations are heterogeneous, inhomogeneous in space, irregular in time, and subject to differing accuracies. In contrast, computational models use knowledge of underlying physics and dynamics to provide a complete description of state evolution in time. Models are also far from perfect: due to model error, uncertainty in the initial conditions and computational limitations, model evolution cannot accurately generate the true state. In order to obtain an analyzed state that is more complete and accurate than the raw observations or model simulations by themselves, data assimilation (henceforth referred to as DA) merges observations into the models.

By its very nature, DA is a complex interdisciplinary subject that involves engineering, geophysics and mathematics. The success of DA depends on the quality of observational technology and geophysical modeling, as well as the scheme that combines them. The DA schemes in use have been built on a variety of mathematical theories and techniques originating in such areas as statistics, dynamical systems and numerical

analysis. Nonetheless, the development of DA has been predominantly led by the geophysical community. Driven by operational demands for numerical weather predictions and thanks to rapidly improving observational technology and computational resources, DA has made momentous practical advances in recent years. As new types of observational data become available at an overwhelming rate and increased knowledge of the underlying geophysical system advances geophysical modeling, the need to develop more sophisticated yet efficient DA schemes grows commensurately.

The goal of this SAMSI program will be to identify outstanding mathematical issues and challenges of geophysical DA while exploring innovative approaches and new directions, in order for geophysical DA to advance further beyond the current state-of-the-art. We will achieve this objective by providing a platform for interdisciplinary collaborations through carefully designed plans for interactive research activities that will span a semester. This will include a workshop series, initiation of collaboration between research groups and individuals, and training of new mathematical scientists (see also ``Workshops and Other Activities" Section below).

The workshop series is designed to achieve the main goal of bringing people with diverse but relevant expertise together. Since the basic concept of DA is not limited to geophysical systems, we will seek the participation of experts in other scientific and engineering fields that share similar needs for estimation and prediction based on incomplete models and diverse data. To cover all anticipated issues and pace ourselves towards the main goal within the limited time frame, we will organize three to four core workshops of medium length % (two, three to five days) in a sequence. To respond to new directions and issues as they emerge during these core workshops, we plan to organize, in addition, spontaneous one-day workshops at SAMSI. This complementary core-spontaneous workshop formula is aimed to effectively stimulating collaboration. To foster collaboration and individual projects further, we will support repeated visits by key people. For the training of new researchers, the first core workshop at the beginning of the semester will have tutorial sessions given by leading experts to cover various aspects of DA.

Because of its interdisciplinary nature, DA is an advanced subject and has been taught only at a handful of universities. A semester-long DA course will follow the tutorial session of the first workshop for new researchers from the SAMSI partner universities. This will form a model for DA courses at other institutions and is anticipated to lead to a basic textbook in the area. Participants of the DA course will be encouraged to take an active part in the spontaneous one-day workshops and develop contacts for their future career.

Advancing toward the main goal of this program can lead to significant contributions to geophysical DA through the development of new statistical, dynamical and computational strategies. SAMSI, as an institute, is uniquely placed to host such an ambitious and timely program.

**Workshops:** Three major workshops are planned:

*1. Kickoff Workshop*, January 23-26, 2005. This workshop will be divided into tutorial sessions (1.5days) and research sessions (1.5days).

2. *Workshop on Mathematical Challenges in Geophysical Data Assimilation,* February 22-26, 2005 at IPAM, UCLA. This will attempt to not only present the current state-of-

art of DA but also bring people together from related fields, such as estimation and control theories, stochastic and dynamical systems.

3. Closing *Workshop* for 1-2 days in Spring, 2005. The aim of this workshop is to gather a small number of people, have key note speakers, and disseminate key findings and directions. It will, in particular, include program officers from funding agencies among invited participants.

**Educational Activities:** Two educational activities will be part of the program:

1. A SAMSI course on Data Assimilation will be taught during Spring, 2005

2. There will be a graduate student Summer school at NCAR, June 12-23, 2005 on "Fusing numerical models and data: Practice to theory to practice." NCAR's new Data Assimilation Research Test Bed (DART) will be used as a teaching aid.

**Program Leaders:** Christopher K.R.T. Jones (Chair) (Dept. of Mathematics, UNC-CH), Kayo Ide (Dept. of Atmospheric Sciences, UCLA), Robert N. Miller (College of Oceanic and Atmospheric Science, OSU), Douglas Nychka (Climate and Global Dynamics Division, NCAR), and Francisco Werner (Dept. of Marine Sciences, UNC-CH).

**Scientific Committee:** Jeffrey Anderson (atmospheric DA), NCAR; Mark Berliner (statistics), Ohio State; Andrew Bennett (ocean DA), Oregon State University; Craig Bishop (atmospheric DA), Navy Research Laboratory Monterey; Montserrat Fuentes (statistics), NC State University; Sujit Ghosh (statistics), NC State University; Eugenia Kalnay (atmospheric DA), University of Maryland; Susan Lozier (ocean data), Duke University; Arthur Mariano (ocean data), University of Miami; Ian McKeague (statistics), Columbia University; Juan Restrepo (atmospheric DA), University of Arizona; Leonard A. Smith (time series prediction), Oxford University, UK; Chris Snyder (atmospheric DA), NCAR; Istvan Szunyogh (atmospheric DA), University of Maryland; Olivier Talagrand (atmospheric DA), Ecole Normal Superier, France; Keith Thompson (ocean DA), Dalhousie University, Canada; Zoltan Toth (atmospheric DA), National Center for Environmental Prediction; Carl Wunsch (Ocean DA), MIT.

**Participants:** Committed and tentatively committed national and international core participants to date are the *SAMSI University Fellow*, Leonard A. Smith (Oxford University, UK), Ian McKeague (Columbia University), Juan Restrepo (Arizona), Marianna Pensky (University of Central Florida), and Roy Chaudhury (University of Central Florida). In addition, we plan to support a number of senior fellows who will actively participate in the SAMSI program by attending spontaneous one-day workshops and visiting SAMSI on return visits for collaboration. Potential senior fellows include Stephen E. Cohn (NASA Goddard Space Flight Center) and Arthur Mariano (U. Miami).

*Postdoctoral Fellows* for the program will be (i) Shree Khare (Atmospheric DA), Princeton University with expected graduation in the summer/fall 2004 (to be officially confirmed) - the position will be jointly support by SAMSI and NCAR; (ii) One postdoctoral fellow, potentially supported for a second year by ONR funding of Jones and Ide on the ocean DA project.

*Faculty Release Fellows* from the partner universities are Sujit Ghosh (Statistics, NCSU), Christopher K.R.T. Jones (Mathematics, UNC-CH), Susan Lozier (Environmental Science, Duke), and Francisco Werner (Marine Sciences, UNC-CH).

*Graduate Fellows* include Livan Liu (Mathematics, UNC), Joe Lucas (Statistics, Duke), and Nichole Mich (Environmental Sciences, Duke).

*Collaborative Organizations:* Potential organizational involvement will come from the Institute for Pure and Applied Mathematics, National Center for Atmospheric Research, National Center for Environmental Prediction, National Center for Integrated and Sustained Ocean Observations, National Research Laboratory, National Aeronautical and Space Administration, Office of Naval Research, and the Global Ocean and Data Assimilation Experiment.

## IV. PLANNING WORKSHOPS

The two planning workshops below will be held in Year 3. As mentioned earlier, we have found planning workshops, often in coordination with NISS or other organizations, to be an extremely effective device, and we anticipate that other planning or hot topic workshops will be held during the year as need arises.

1. A workshop *The Design and Analysis of Computer Experiments for Complex Systems* will be held at the Banff Conference Center from 7/13/04-7/17/04. The workshop is being organized by Derek Bingham (Department of Statistics an Actuarial Sciences, Simon Fraser University) and Jim Berger (SAMSI). It is a joint workshop of the Canadian National Program for Complex Data Structures, SAMSI, and NISS. A major focus of the workshop is to consider the possibility and nature of a future SAMSI program in the area, since so many SAMSI programs involve the interface of statistics, mathematics and computer modeling.

2. Leaders and potential participants of the likely Year 4 Program *Financial Mathematics and Econometrics* will meet in early June, 2005, at a joint workshop of SAMSI and the Centre de Recherches Mathématiques in Montreal, to determine the key foci, both theoretical and applied, of the program.

# B. Scientific Themes for Later Years

The programs listed below have not yet been formally approved, but all are well along in the development cycle and we are confident that they will be approved and implemented.

I. FINANCIAL MATHEMATICS AND ECONOMETRICS (tentative for Fall, 2005)

Financial Mathematics and Econometrics in the broad sense, from Mathematical Finance to Financial Engineering and Econometric implementation, is a rapidly expanding and growing field. It consists of multi-disciplinary and overlapping set of fields which involves disciplines such as: Applied Mathematics, Economics and Finance, Econometrics and Statistics. Since the introduction of the geometric Brownian motion as a tool for modeling stock price evolution, and the discovery in the early seventies of the Black-Scholes formula for pricing options, a lot has been achieved. A tremendous amount of work has sought to understand and explain option prices observed in the markets, and to build tools to handle the associated risks. Needless to say derivatives markets currently represent an important percentage of trades and investments. Moreover, in the last decades four Nobel prizes in Economics were attributed to research in the fields of Financial Mathematics (Merton and Scholes in 1997) and Financial Econometrics (Engle and Granger in 2003).

The proposed goal of a SAMSI program in Financial Mathematics and Econometrics would be to bring together these disciplines, and open a discussion on what is really important and what is missing in the three essential tasks: {\bf modeling, handling data and computing}, in domains ranging from financial and energy derivatives to real options.

*Modeling.* In equity markets there is a profusion of models ranging from local volatility to stochastic volatility, multi factors with and without jumps, based on Brownian motions or Levy processes. The situation is similar in fixed income markets with short rates models, HJM or BGM models to name only a few. There is also a variety of discrete time models as the ARCH family for instance. The fundamental question of relevance of these models will be addressed in the program as well as links between physical measures and pricing measures through market prices of risks. Closely connected are the problems of hedging and portfolio optimization which will also be addressed.

*Data.* The size of financial data can be considerable when looking at high frequency data for large numbers of stocks for instance. Data is essential in the modeling part in at least two ways: writing models which capture the main effects seen in the data (for instance 'are jumps present?') and calibrating the models with an assessment of the stability of the parameters. A lot has been done in this direction in Statistics and Econometrics and to a lesser extent in Applied Mathematics. The program will bring these disciplines together, present the state of the art and discuss issues on choosing, preparing and using financial data. The program will make sure that statistical software companies are involved.

*Computation.* Once a model has been written and calibrated to data, it remains to compute quantities of interest. For instance, in option pricing, one has to compute expected values along the trajectories (time evolution) of multidimensional stochastic

processes. These quantities are also often obtained as solutions of partial differential equations (or inequalities) with various boundary conditions. The program will address the question of choosing the most efficient computational method for classes of problems. In particular Monte Carlo methods and numerical methods will be discussed, keeping in mind that the computational difficulty has a feedback effect on the modeling and data calibrating parts.

Activities of the program will include opening and closing workshops, regular working group meetings, and two courses, 'Computational Genomics for Infectious Disease' and 'Population Dynamics for Infectious Disease.'

There are a wide variety of potential industrial and governmental organizations that are likely to seek involvement in the program. In addition, the NIH, National Institute of Allergy and Infectious Disease, Burroughs-Wellcome Foundation, Ellis Foundation, and the Gates Foundation are potentially interested in the program and are potential sources of additional funding.

We expect this program to have an impact on the practice of computational biology, mathematics and statistics, by making biologists aware of the value to be gained by involving mathematically sophisticated personnel in their research teams, and by making statisticians and mathematicians aware of the most outstanding problems in biology.

*Potential Program Leaders:* Marco Avellaneda (NYU, Mathematics), Jean-Pierre Fouque} (NC State, Mathematics), Eric Ghysels (UNC, Economics), Ronnie Sircar (Princeton, Operations Research \& Financial Engineering), and Ruey Tsay (University of Chicago, Graduate School of Business Statistics department).

*Potential Scientific Committee:* A group of internationally distinguished scholars have accepted to be members of this committee. They are Ole E. Barndorff-Nielsen (Centre for Mathematical Physics and Stochastics, Aarhus, Denmark), Rene Carmona (Princeton University), Darrell Duffie} (Stanford University), Lars Hansen (University of Chicago), and Robert Jarrow (Cornell University).

## II. HIGH DIMENSIONAL INFERENCE AND RANDOM MATRICES (tentative for Fall, 2005)

The recent growth of interest in 'random matrix theory' (RMT) in many areas of mathematics suggests possible developments in statistics, and the idea of SAMSI holding a semester program. As one example, RMT seems to offer tools that allow one to revisit classical multivariate statistics from a 'large n, large p' perspective, and perhaps derive approximations and insights that were not available due to the complexity of the distribution theory for fixed n,p or fixed p, large n.

If one takes real statistical contexts and abstracts mathematical/probabilistic questions, one is very likely to come up with issues that have not yet been addressed by the RMT community, and yet are accessible with current tools. For example, the distribution of extreme Wishart eigenvalues for non-identity covariance matrix is a rather natural statistics question that had not been looked at in RMT, and yet results are now beginning to emerge.

A planning meeting was held at the American Institute of Mathematics, Palo Alto, February 29 and April 1. The participants and general structure of the meeting was described in Section I.E.5 above. Major research areas, with important subtopics were agreed upon:
*Extreme Sample Eigenvalues* (Large *n, p* Setting): PCA, CCA; distributions under alternative hypotheses; Integrable systems - long term project; beta = 1; Connection to financial math program.
*Sample Eigenvectors:* Consistency and distribution; Smoothing, filtering of lead estimated eigenvectors.
*Empirical Distribution Of Eigenvalues* (Global): Marcenko-Pastur applications; contiguity; CLT for linear statistics.
*Approximations Of Empirical Data:* Compact energy decay - finite range approximations; Invariant approximations; Random Schroedinger Operators
*Design Of Snapshots:* Design measure; Determininstic errors (treated as stochastic?); PDE induced smoothing of Principal Orthogonal Decomposition (POD) empirical functions; Prespecified basis functions.
*Nonlinear / Topological Approaches To Dimensional Reduction:* Data sets; Dynamical systems; Uncertainty assessment.
*Bayesian Version Of Regularization In Large P Problems / Invariance Structure:* Coherence with frequentist uncertainty assessment; Prior distributions on covariance structures: dispensing with Vandermondes; Banded random matrices.
*Stochastic Evolution Of Random Matrices:* Dyson Brownian Motion model; Wishart Processes.
*Statistical Issues In Empirical Orthogonal Functions (EOFs) In Climatology*

*Potential Program Participants were listed:* Ben Santer, Myles Allen, Gabi Herguel, Harold Widom, Jean Bernard Zuber, Don Richards, Mark Adler, P. van Moerbeke, Alexander Its, Percy Deift, Stephanos Venakides, Thomas Guhr, Jianqing Fan, Piet Groeneboom, Friedrich Götze, Brenda MacGibbon, Alan Edelman, Peter Forrester, Ofer Zeitouni, Persi Diaconis, Dave Donoho, Amir Dembo, Steve Evans, Zidong Bai, Sasha Soshnikov, Eric Rains, Neil O'Connell, Kurt Johansson, Peter Miller, Edris Titi, Ioannis Kevrekedis, Mikhail Belkin, Bas Klein, Jayanta Ghosh, Jon Wellner, Alexander Tsybakov, Vladimir Koltchinskii, Sara van de Geer, Aad van de Vaart, Evarist Giné, Jonathan Taylor, Estelle Basor, Kelly Wieand, Tom Spencer, Josh Tenenbaum, Lawrence Saul Kesnel.

*Possible Leaders of Program:* Iain M. Johnstone (Statistics, Stanford), Craig A. Tracy (Mathematics, U. C. Davis), Ken McLaughlin (Mathematics, UNC)

## III. ASTROSTATISTICS (tentative for Spring, 2005)

A vast range of statistical problems arise in modern astronomical and space sciences research, particularly due to the flood of data produced by space-based astronomical surveys at many wave-bands. A resurgence of interest in statistical methods has emerged among space scientists as they seek insights into the physical phenomena underlying such complex data. Researchers at the frontiers confront problems for which current statistical approaches either inadequately utilize known methods or require the development of new

methods. In contrast with the biological and social sciences, the statistical needs of physical scientists have been neglected during recent decades.

To cope with the current and future needs of astronomy missions require concerted efforts by cross-disciplinary collaborations involving astronomers, computer scientists, mathematicians and statisticians. SAMSI is an ideal place from which to broadcast these issues and involve the wider statistical and applied mathematical communities. As I am familiar with and in constant touch with many of the cross-disciplinary groups working in astrostatistics, I can bring in research groups focused on different sub areas to SAMSI.

I propose a semester long Astrostatistics program at SAMSI in Spring 2006. A vital ingredient of the planned astrostatistics program at SAMSI is to provide a single geographical location - a crossroads - where researchers at the interface between statistics, applied mathematics, astronomy, and particle physics can congregate and initiate lasting collaborations. The participation by graduate students and postdocs give them a rare opportunity to develop skills needed for cross-disciplinary work.

Astronomy at the beginning of the 21st century, and particularly research arising from robotic space-based observatories, finds itself with serious challenges in statistical treatments of data to achieve its astrophysical goals. Innumerable issues arise in the scientific interpretation of astronomical studies. Some issues involve sampling, multivariate and survival analysis, while others involve image and spatial analysis, signal processing or time series analysis. Nonlinear regression is needed to model the spectra of astronomical objects in terms of continuum and line components deriving from the quantum mechanical properties of matter. Here are a few of the questions that arise:

Is a collection of objects chosen for study an unbiased sample of the vast underlying population? When should a collection of objects be divided into two or more classes? What is the intrinsic relationship between two properties of a class, particularly in the presence of confounding variables such as redshift? How can we answer such questions in the presence of flux-limited samples and flux-dependent error bars? When is a blip in a spectrum or image a real signal rather than noise? How do we characterize blips embedded in larger structures? When is a signal variable rather than constant? How do we characterize the vast range of periodic, correlated and stochastic variations ranging from the Doppler wobble of normal stars due to invisible planets, X-ray manifestations of accretion onto black holes, and gamma-ray bursts from the exotic end-states of stellar evolution? How do we understand the 3-to-6-dimensional spatial point processes representing the location and motions of stars in the Galaxy or Galaxies in the Universe? How do we understand the structure of continuous entities like the cosmic microwave background or the interstellar medium?

The typical quality of statistical methodology used to address such complex questions has not been very high among astronomers. Most astronomers are largely unaware of a host of important statistical and computational developments of the past half century: robust methods, bootstrap resampling, hidden Markov models, empirical Bayes and James-Stein estimation, survival analysis, semi-parametric methods, Bayesian decision theory, and much more. Some methods lie on the interface between mathematical statistics, mathematics, and computational methods: consider, for example, the EM Algorithm, Kalman filter and Monte Carlo Markov chain for likelihood calculations.

Other statistical issues do not appear in research journals but rather arise deep inside the complex machinery of modern astronomical observatories. Many testing,

monitoring, compressing, fitting and even intelligent decision-making operations are embedded in the operation, calibration and data reduction process of a contemporary astronomical satellite. With advances in high-speed radiation-hardened chips and high-data rate detectors, sophisticated data analysis operations often take place on-board. Telemetered data are then subject to pipeline processing which provide the basic input to hundreds of astronomical studies. Most of the codes are developed by engineers and scientists who have little formal training in statistics or applied mathematics.

*Potential program leaders*, all of whom are participating in planning discussions, include: G. J. Babu (chair – statistics, Pennsylvania State), Eric Feigelson (astronomy, Penn State), Donald Richards (Statistics, Virginia), Alanna Connors (Astronomy, ??) and Larry Wasserman (Statistics, Carnegie-Mellon). Other scientists that will be consulted about the SAMSI program include Françoise Genova (Centre des Données Astronomiques de Strasbourg), George Djorgovski (Caltech, Astronomy), Iian Johnstone (Stanford, statistics), Fionn Murtagh (Queen's, Computer Science), John Rice (Berkeley, Statistics), David van Dyk (UC Irvine, statistics), Lee Samuel Finn (Penn State, physics), Ajit Kembhavi (IUUCA, astronomy), Thomas Loredo (Cornell, astronomy), Louis Lyons (Oxford, Physics), Jean-Luc Starck (Saclay, astronomy), Vicent Martinez (Valencia, astronomy), and Philip Stark (statistics, Berkeley).

IV. NATIONAL SECURITY AND HOMELAND DEFENSE (Possible for Year 4)
A program in this area is under strong consideration, given the highly successful outcome of the hot topics workshop in the area (see Section I.E.5).

# **APPENDIX A –**Follow-up for Program on Inverse Problem Methodology in Complex Stochastic Models

While it is difficult to quantify all the benefits and results of the IP Program after one year, there are some clear metrics that can be used to evaluate its impact. Since the primary benefits should be expected for participants who are at early stages in their careers, we summarize first the tangibles perceived for student and postdoc participants.

SUMMARY OF SAMSI POSTDOCS' SUBSEQUENT ACTIVITIES:

There were three SAMSI funded postdocs in the IP Program: John Bardsley, Yanyuan Ma, and Danny Walsh. All three participated heavily in (indeed, helped lead) the June, 2003 Undergraduate Workshop on Inverse Problems. They have followed somewhat different courses since July, 2003.

1. After the year at SAMSI, Bardsley accepted a position as Assistant Professor of Mathematics at the University of Montana, Missoula, MT. He has remained active in research, producing three research papers [1], [2], [3], one ([2]) of which combines applied mathematical and statistical techniques in the SAMSI spirit to address electromagnetic interrogation problems.
2. Yanyuan Ma came to SAMSI with a Ph.D. in applied mathematics with the determined intention to become equally expert in statistical methods. Indeed, her goal was to become trained so as to follow a career in statistics, eventually obtaining a faculty position in a stat department at a Research I university. After the first year at SAMSI, she spent her 2nd year as a CRSC postdoc, working with statisticians (resulting in publications [4}-[8]) to improve her statistical contributions, while contributing to the HIV project (described below) at CRSC led by Banks and Davidian. She has been very successful in both her contributions and her career objectives. She has accepted a position beginning Fall, 2004, as Assistant Professor of Statistics at Texas A&M University.
3. After his year as a SAMSI postdoc in the IP Program, Walsh joined NISS as a postdoc, working on a model validation project. He has accepted a position in Statistics at Massey University, Auckland, NZ, beginning Fall semester, 2004.

SUMMARY OF SAMSI GRADUATE STUDENTS' SUBSEQUENT ACTIVITIES

A number of graduate students participated in the IP Program. SAMSI Graduate Fellows included Karen Chiswell and Jeff Hood. SAMSI Graduate Associates included Brian Adams, Brandy Benedict and Nathan Gibson. They all participated in the working sessions, the IP course, and in the Undergrad Workshop in June, 2003. While they are still involved in completing their degree requirements, their research programs have been significantly impacted by their participation in the SAMSI IP Program. Karen Chiswell is pursuing a thesis on PBPK modeling with strong applied math and statistical components; both Banks and Davidian are serving on her graduate committee. Jeff Hood contributed efforts to organizing the notes from the IP course (taught by Banks and

Davidian) after the course was completed. He is now working on a thesis that combines statistical modeling and probabilistic molecular formulations with dynamical systems for reptation models of polymeric materials (viscoelastic fluids, rubbers, biotissues, etc.). Adams is playing a major role in a long term, NIGMS funded research program on HIV modeling that entails both applied math and statistical methodology. He contributed significant efforts to several papers ([10],[16]) and his thesis on modeling of HIV will embody clear evidence of the SAMSI applied math/statistics synthesis. Gibson is working on problems in the other major research theme of the SAMSI IP Program, electromagnetics. The October, 2002, Electromagnetics Day, held at SAMSI as an IP activity, led to several interesting collaborations, including a NASA sponsored project on E&M interrogation of the shuttle foam for damages that is the focus of Gibson's thesis. In this effort he has used statistical methods to give uncertainty bounds on the estimation algorithms used to detect damage as reported in [11] as well as probabilistic modeling of polarization in complex materials [14].

IMPACT ON "NEW RESEARCHERS"

A number of new researchers who visited SAMSI several times during the IP Program on inverse problems have visibly benefited from their participation. Among these are A. Ackleh (Univ. Louisiana, Lafayette), J. Banks (Univ. Washington, Tacoma), C. Cole (Meredith College), S. Ghosh (NCSU), G. Pinter (Univ. Wisconsin, Milwaukee) and C. Wikle (Univ. Missouri). Ackleh is now leading his group in combining statistical and applied math methods in the area of biological models ( [17], [18] ) and has collaborated with C. Cole on some of this work. J. Banks, a biologist, has collaborated with Adams ( [10] )on dynamical and statistical methods for inverse problems in pesticide management of insects, while Ghosh has become involved in the stat/applied math HIV modeling project at NCSU, one component of which combines Bayesian methods with a Prohorov framework for estimation of parameters. Pinter, after extensive participation in the IP Program, has become very active in using probabilistic methods to model dynamical systems arising in materials ([12], [13], [15] ). Wikle, a statistician who has been an extensive visitor to SAMSI after initial participation in the IP Program, has recently begun collaborations with applied mathematicians on modeling of migratory patterns of of birds.

IMPACT ON "SENIOR RESEARCHERS"

Among senior researchers (along with their students) who have clearly been affected by the SAMSI IP Program are H.T. Banks (NCSU), D. Cioranescu (U. Paris VI), and J. Whiteman and S. Shaw (Brunel University). Cioranescu and her students have become involved in comparing homogenization methods with probabilistic averaging in inverse problems in electromagnetics while Whiteman and Shaw have begun collaborations with the authors of [12], [13], [15] on probabilistic methods for viscoelastic system inverse problems.

PUBLICATIONS: (that resulted directly from interactions and collaborations stimulated by the SAMSI IP Program)

[1] H.T. Banks and J. Bardsley, Well-posedness for systems arising in time domain electromagnetics in dielectrics, CRSC-TR03-27, July, 2003; *Communications in Applied Analysis*, to appear.

[2] H.T. Banks and J. Bardsley, Parameter identification for a dispersive dielectric in 2D electromagnetics: Forward and inverse methodology with statistical considerations, CRSC-TR03-48, December, 2003*; Inverse Problems,* submitted.

[3] J. Bardsley, A bound-constrained Levenberg-Marquardt algorithm for a parameter identification problem in electromagnetics*, J. Inverse and Ill-posed Problems,* submitted.

[4] Y. Ma and M.G. Genton, A semiparametric class of generalized skew-elliptical distributions, *Scandinavian Journal of Statistics,* to appear.

[5] A.A. Tsiatis and Y. Ma, Locally efficient semiparametric estimators for functional measurement error models, *Biometrika,* to appear.

[6] Y. Ma, M.G. Genton and M. Davidian, Linear mixed effect models with semiparametric generalized skew-elliptical random effects, in``*Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality",* Genton, M. G. (editor), Chapman \& Hall / CRC, 2004, to appear.

[7] Y. Ma, M.G. Genton and A.A. Tsiatis, Locally efficient semiparametric estimators for generalized skew-elliptical distributions, *J. American Statistical Association,* submitted.

[8] Y. Ma and A.A. Tsiatis, Closed form semiparametric estimators for mixed models with complete sufficient statistic, to be submitted.

[9] H.T. Banks, Y. Ma and L.K. Potter, A simulation based comparison between parametric and semiparametric methods in a PBPK model, *J. Inverse and Ill-posed Problems,* to be submitted.

[10] B.M. Adams, J.E. Banks, H.T. Banks and J.D. Stark, Population dynamics models in plant-insect-herbivore-pesticide interactions, CRSC-TR03-12, March, 2003, Revised, August, 2003; *Math. Biosci.,* submitted.

[11] H.T. Banks, N.L. Gibson and W.P. Winfree, Electromagnetic crack detection inverse problems using Terhertz interrogating signals, CRSC-TR03-40, October, 2003; *Inverse Problems*, submitted.

[12] H.T. Banks, N.G. Medhin and G.A. Pinter, Multiscale considerations in modeling of nonlinear elastomers, CRSC-TR03-42,October, 2003; *J. Comp. Meth. Sci. and Engr.,* to appear.

[13] H.T. Banks, N.G. Medhin and G.A. Pinter, Nonlinear reptation in molecular based hysteresis models for polymers, CRSC-TR03-45, December, 2003; *Quarterly Applied Math.,* to appear.

[14] H.T. Banks and N.L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, CRSC-TR04-01, January, 2004; *Applied Math. Letters*, submitted.

[15] H.T. Banks and G.A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue , CRSC-TR04-03, January, 2004; *SIAM J. Multiscale Modeling and Simulation,* submitted.

[16] B.M. Adams, H.T. Banks, M. Davidian, H.D. Kwon, H.T. Tran, S.N. Wynne and E.S. Rosenberg , HIV dynamics: Modeling, data analysis, and optimal treatment protocols, CRSC-TR04-05, February, 2004; *J. Comp. and Appl. Math.,* submitted.

[17] A.S. Ackleh, K. Deng, C. Cole and H.T. Tran, Existence-uniqueness and monotone approximation for an erythropoiesis age-structured model, CRSC-TR03-21, April, 2003; *J. Math. Anal. Appl.,* submitted.

[18] A.S. Ackleh, H.T. Banks, K. Deng, S. Hu, Parameter estimation in a coupled system of nonlinear size-structured populations, to be submitted.

## APPENDIX D – Workshop Participants

TWO tables of participants for most of the SAMSI workshops follow. The first table lists only the individuals who received support. The second table lists all workshop participants. The minority status of each participant is available, but we do not include the information here because of privacy issues; the summaries in Section I.H. were compiled from this data.

The key to **Status** entry is as follows:

NRG – New Researcher or Graduate Student
FP – Faculty/Professional
S – Student (Education & Outreach)
F – Faculty (Education & Outreach)

**Inverse Problems Methodology in Complex Stochastic Problems**
**Closing Workshop**
*Supported Workshop Participants*
NISS-SAMSI Building
May 14-15, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Ackleh, Azmy | M | U of Lousiana | Mathematics | FP |
| Engl, Heinz | M | Kepler U | Mathematics | FP |
| Pinter, Gabrielle | F | U of Wisconsin | Mathematics | FP |
| Rosner, Gary | M | MD Anderson Cancer Research Center | Biostatistics | FP |
| Sheiner, Lewis | M | U of California-San Francisco | Statistics | FP |
| Somersalo, Erkki | M | Helsinki U of Technology | Mathematics | FP |
| Wang, Chunming | M | U of Southern California | Mathematics | FP |
| Wikle, Chris | M | U of Missouri | Statistics | FP |

**Inverse Problems Methodology in Complex Stochastic Problems**
**Closing Workshop**
*Workshop Participants*
NISS-SAMSI Building
May 14-15, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Ackleh, Azmy | M | U of Lousiana | Mathematics | FP |
| Adams, Brian | M | North Carolina State U | Mathematics | NRG |
| Bardsley, Johnathan | M | North Carolina State U & SAMSI | Mathematics | NRG |
| Benedict, Brandy | F | North Carolina State U | Mathematics | NRG |
| Berger, Jim | M | SAMSI | Statistics | FP |
| Chiswell, Karen | F | North Carolina State U | Statistics | NRG |
| Cole, Cammey | F | Meredith College | Mathematics | FP |
| Davidian, Marie | F | North Carolina State U | Statistics | FP |
| Engl, Heinz | M | Kepler U | Mathematics | FP |
| Gibson, Nathan | M | North Carolina State U | Mathematics | NRG |
| Hood, Jeffrey | M | North Carolina State U | Mathematics | NRG |
| Hurtado, Gerardo | M | SAS Institute | | FP |
| Joyner, Michele | F | State U of West Georgia | Mathematics | FP |
| Kepler, Grace | F | North Carolina State U | Mathematics | NRG |
| Ma, Yanyuan | F | North Carolina State U | Mathematics | NRG |
| Medhim, Negash | M | North Carolina State U | Mathematics | FP |
| Pinter, Gabrielle | F | U of Wisconsin | Mathematics | FP |

| | | | | |
|---|---|---|---|---|
| Potter, Laura | F | GlaxoSmithKline | | FP |
| Rosner, Gary | M | MD Anderson Cancer Research Center | Biostatistics | FP |
| Sheiner, Lewis | M | U of California-San Francisco | Statistics | FP |
| Somersalo, Erkki | M | Helsinki U of Technology | Mathematics | FP |
| Tran, Hien | M | North Carolina State U | Mathematics | FP |
| Walsh, Daniel | M | SAMSI | | NRG |
| Wang, Chunming | M | U of Southern California | Mathematics | FP |
| Wikle, Chris | M | U of Missouri | Statistics | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |

**Large-Scale Computer Models for Environmental Systems**
**One-Day Workshop on Porous Media Processes**
*Supported Workshop Participants*
NISS-SAMSI Building
May 16, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Wildenschild, Dorothe | F | Oregon State U | Geosciences | FP |

**Large-Scale Computer Models for Environmental Systems**
**One-Day Workshop on Porous Media Processes**
*Workshop Participants*
NISS-SAMSI Building
May 16, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Abhishek, Chandra | M | U of North Carolina | Environmental Eng | NRG |
| Adalsteinsson, David | M | U of North Carolina | Mathematics | FP |

| | | | | |
|---|---|---|---|---|
| Arellano, Ave | M | Duke U | Environment | NRG |
| Camassa, Roberto | M | U of North Carolina | Mathematics | FP |
| Chen, Li | F | North Carolina State U | Statistics | NRG |
| Culligan, Katherine | F | U of Notre Dame | Civil Engineering | NRG |
| Davis, Jerry | M | North Carolina State U | Meteorology | FP |
| Dennis, Robin | M | U.S. EPA | Atmos. Modeling | FP |
| Farthing, Matthew | M | U of North Carolina | Environmental Eng | NRG |
| Finkelstein, Peter | M | U.S. EPA | | FP |
| Forest, Greg | M | U of North Carolina | Mathematics | FP |
| Fuentes, Montserrat | F | North Carolina State U | Statistics | FP |
| Genton, Marc | M | North Carolina State U | Statistics | FP |
| Grady, Amy | F | NISS | | FP |
| Gray, William | M | U of Notre Dame | Civil Engineering | FP |
| Holland, Dave | M | U.S. EPA | | FP |
| Hilpert, Markus | M | Johns Hopkins U | Geography & Environmental Eng | FP |
| Hurd, Harry | M | U of North Carolina | Statistics | FP |
| Irwin, John | M | U.S. EPA | | FP |
| Johnson, Deona | F | U of North Carolina | Environmental Eng | NRG |
| Jones, Chris | M | U of North Carolina | Mathematics | FP |
| Kolenikov, Stas | M | U of North Carolina | Statistics | NRG |

| | | | | |
|---|---|---|---|---|
| Kuznetsov, Leonid | M | U of North Carolina | Mathematics | FP |
| Lee, Long | M | U of North Carolina | Mathematics | NRG |
| Li, Huina | M | U of North Carolina | Environmental Eng | NRG |
| Lopes, Brian | M | U of North Carolina | Statistics | NRG |
| McClure, James | M | U of North Carolina | Environmental Eng | NRG |
| McLaughlin, Richard | M | U of North Carolina | Mathematics | FP |
| Miller, Cass | M | U of North Carolina | Environmental Eng | FP |
| Nail, Amy | F | North Carolina State U | Statistics | NRG |
| Natvig, Bent | M | U of Oslo | Mathematics | FP |
| Nolte, David | M | Purdue U | Physics | FP |
| Pan, Doris | F | U of North Carolina | Environmental Eng | NRG |
| Pyrak-Nolte, Laura | F | Purdue U | Physics | FP |
| Reese, Jill | F | North Carolina State U | Mathematics | NRG |
| Sahu, Sujit | M | U of Southampton | Mathematics | FP |
| Scotti, Alberto | M | U of North Carolina | Marine Sciences | FP |
| Serre, Marc | M | U of North Carolina | Environmental Eng | FP |
| Smith, Richard | M | U of North Carolina | Statistics | FP |
| Tvete, Ingunn Fride | F | U of Oslo | Mathematics | NRG |
| Wildenschild, Dorothe | F | Oregon State U | Geosciences | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |
| Yu, Hwa-Lung | M | U of North Carolina | Environmental Eng | NRG |

| | | | | |
|---|---|---|---|---|
| Zhu, Zhengyuan | M | U of North Carolina | Statistics | NRG |
| Zidek, James | M | U of British Columbia | Statistics | FP |

**Large-Scale Computer Models for Environmental Systems**
**Workshop on Spatio-Temporal Modeling**
NCAR—Boulder, CO
*Supported Workshop Participants*
June 1-6, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Berliner, Mark | M | Ohio State U | Statistics | FP |
| Calder, Catherine | F | Duke U | Statistics | NRG |
| Caragea, Petruta | F | U of North Carolina | Statistics | NRG |
| Carlin, Brad | M | U of Minnesota | Biostatistics | FP |
| Chen, Li | F | North Carolina State U | Statistics | NRG |
| Christakos, George | M | U of North Carolina | Engineering & Environmental Sci | FP |
| Craigmile, Peter | M | Ohio State U | Statistics | FP |
| Cressie, Noel | M | Ohio State U | Statistics | FP |
| Davis, Jerry | M | North Carolina State U | Marine Sciences | FP |
| Dominici, Francesca | F | John Hopkins U | Biostatistics | FP |
| Drignei, Dorin | | Iowa State U | Statistics | NRG |
| Ferreira, Marco | M | Federal U Rio Janeiro | Statistics | FP |
| Fink, Daniel | M | Cornell U | Social Statistics | FP |
| Fuentes, Montserrat | F | North Carolina State U | Statistics | FP |

| | | | | |
|---|---|---|---|---|
| Gelfand, Alan | M | Duke U | Statistics | FP |
| Grady, Amy | F | NISS | | FP |
| Guillas, Serge | M | U of Chicago | Statistics | NRG |
| Hegerl, Gabriele | F | Duke U | Environment | FP |
| Holt, John | M | U of Guelph | Mathematics & Statistics | FP |
| Hooten, Mevin | | U of Missouri | Statistics | NRG |
| Huerta, Gabriel | M | U of New Mexico | Statistics | FP |
| Jun, Mikyoung | F | U of Chicago | Statistics | NRG |
| Kleeman, Richard | M | New York U | Courant Institute | FP |
| Kolenikov, Stanislav | F | U of North Carolina | Statistics | NRG |
| Lam, Eric | M | Ohio State U | Statistics | NRG |
| Lee, Jaechoul | M | U of Georgia | Statistics | NRG |
| Lopes, Brian | M | U of North Carolina | Statistics | NRG |
| Majda, Andrew | M | New York U | Courant Institute | FP |
| Majumdar, Ananda | F | U of Connecticut | Statistics | NRG |
| McLaughlin, Richard | M | U of North Carolina | Mathematics | FP |
| Nail, Amy | F | North Carolina State U | Statistics | NRG |
| Ozaksoy, Isin | | North Carolina State U | Statistics | NRG |
| Sanso, Bruno | M | U of California-Santa Cruz | Applied Mathematics & Statistics | FP |
| Santer, Ben | M | Lawrence Livermore National Laboratory | | FP |
| Serre, Marc | M | U of North Carolina | Engineering & Environmental Sci | FP |

| | | | | |
|---|---|---|---|---|
| Smith, Richard | M | U of North Carolina | Statistics | FP |
| Stein, Michael | M | U of Chicago | Statistics | FP |
| Wikle, Chris | M | U of Missouri | Statistics | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |
| Zhu, Jun | | U of Wisconsin | Statistics | FP |
| Zhu, Zhengyuan | M | U of North Carolina | Statistics | NRG |

**Large-Scale Computer Models for Environmental Systems**
**Workshop on Spatio-Temporal Modeling**
NCAR—Boulder, CO
*Workshop Participants*
June 1-6, 2003

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Anderes, Ethan | M | U of Chicago | Statistics | NRG |
| Anderson, Jeffery | M | UCAR | Climate & Global Dynamics | FP |
| Arab, Ali | | U of Missouri | Statistics | NRG |
| Berger, James | M | SAMSI & Duke U | Statistics | FP |
| Berger, Roger | M | NSF | | FP |
| Berliner, Mark | M | Ohio State U | Statistics | FP |
| Breidt, Jay | M | Colorado State U | Statistics | FP |
| Calder, Catherine | F | Duke U | Statistics | NRG |
| Caragea, Petruta | F | U of North Carolina | Statistics | NRG |
| Carlin, Brad | M | U of Minnesota | Biostatistics | FP |

| | | | | |
|---|---|---|---|---|
| Carroll, Steven | M | Oregon State U | Statistics | FP |
| Chen, Li | F | North Carolina State U | Statistics | NRG |
| Christ, Aaron | M | U of Iowa | Statistics | NRG |
| Christakos, George | M | U of North Carolina | Engineering & Environmental Sci | FP |
| Coar, William | M | Colorado State U | Statistics | NRG |
| Cooley, Dan | M | U of Colorado | Applied Mathematics | NRG |
| Craigmile, Peter | M | Ohio State U | Statistics | FP |
| Cressie, Noel | M | Ohio State U | Statistics | FP |
| Davis, Jerry | M | North Carolina State U | Marine Sciences | FP |
| Davis, Richard | M | Colorado State U | Statistics | FP |
| Dennis, Robin | M | U.S. EPA | | FP |
| Dominici, Francesca | F | John Hopkins U | Biostatistics | FP |
| Draghicescu, Dana | M | U of Chicago | Statistics | NRG |
| Drignei, Dorin | | Iowa State U | Statistics | NRG |
| Dunsmuir, William | M | U of New South Wales | Statistics | FP |
| Eubank, Randall | M | Texas A&M U | Statistics | FP |
| Farnsworth, Matt | M | Colorado State U | Ecology | NRG |
| Ferreira, Marco | M | Federal U Rio Janeiro | Statistics | FP |
| Fink, Daniel | M | Cornell U | Social Statistics | FP |
| Fuentes, Montserrat | F | North Carolina State U | Statistics | FP |
| Gel, Yulia | F | U of Washington | Statistics | NRG |

| | | | | |
|---|---|---|---|---|
| Gelfand, Alan | M | Duke U | Statistics | FP |
| Genton, Marc | M | North Carolina State U | Statistics | FP |
| Grady, Amy | F | NISS | | FP |
| Guillas, Serge | M | U of Chicago | Statistics | NRG |
| Haven, Kyle | M | New York U | Courant Institute | NRG |
| Hegerl, Gabriele | F | Duke U | Environment | FP |
| Holt, John | M | U of Guelph | Mathematics & Statistics | FP |
| Hooten, Mevin | | U of Missouri | Statistics | NRG |
| Huerta, Gabriel | M | U of New Mexico | Statistics | FP |
| Huzurbazar, Snehalata | F | U of Wyoming | Statistics | FP |
| Johnson, Devin | M | Colorado State U | Statistics | NRG |
| Jun, Mikyoung | F | U of Chicago | Statistics | NRG |
| Kleeman, Richard | M | New York U | Courant Institute | FP |
| Kolenikov, Stanislav | F | U of North Carolina | Statistics | NRG |
| Lam, Eric | M | Ohio State U | Statistics | NRG |
| Lee, Jaechoul | M | U of Georgia | Statistics | NRG |
| Lopes, Brian | M | U of North Carolina | Statistics | NRG |
| Madsen, Kristen | F | North Carolina State U | Mathematics | NRG |
| Majda, Andrew | M | New York U | Courant Institute | FP |
| Majumdar, Ananda | F | U of Connecticut | Statistics | NRG |
| McBride, Sandra | F | Duke U | Statistics | NRG |

| | | | | |
|---|---|---|---|---|
| McLaughlin, Richard | M | U of North Carolina | Mathematics | FP |
| Merton, Andrew | M | Colorado State U | Statistics | NRG |
| Nail, Amy | F | North Carolina State U | Statistics | NRG |
| Nychka, Doug | M | NCAR | Climate & Global Dynamics | FP |
| Ozaksoy, Isin | | North Carolina State U | Statistics | NRG |
| Paciorek, Christopher | M | Carnegie Mellon U | Statistics | NRG |
| Sanso, Bruno | M | U of California-Santa Cruz | Applied Mathematics & Statistics | FP |
| Santer, Ben | M | Lawrence Livermore National Laboratory | | FP |
| Schaffrin, Burkhard | M | Ohio State U | Civil & Environmental Engineering | FP |
| Scholzel, Christian | M | U of Bonn | Meterology | NRG |
| Serre, Marc | M | U of North Carolina | Engineering & Environmental Sci | FP |
| Smith, Richard | M | U of North Carolina | Statistics | FP |
| Stein, Michael | M | U of Chicago | Statistics | FP |
| Stroud, Jonathan | M | U of Pennsylvania | Statistics | NRG |
| Urquhart, Scott | M | Colorado State U | Statistics | FP |
| Wigley, Tom | M | NCAR | Climate & Global Dynamics | FP |
| Wikle, Chris | M | U of Missouri | Statistics | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |
| Yongku, Kim | | Ohio State U | Statistics | NRG |
| Zhu, Jun | | U of Wisconsin | Statistics | FP |
| Zhu, Zhengyuan | M | U of North Carolina | Statistics | NRG |

**Education & Outreach Program**
**SAMSI/CRSC Interdisciplinary Workshop for Undergraduates**
Campus of North Carolina State University
*Supported Workshop Participants*
June 9-13, 2003

| Name | Gender | Affiliation | Major/Department | Status |
|---|---|---|---|---|
| Billing, Emily | F | U of North Carolina-Wilmington | Mathematics and Physics | S |
| Bakewell, Edward | M | U of North Carolina-Wilmington | Mathematics & Statistics | S |
| Charpentier, Heidi | F | U of North Carolina-Asheville | Applied Mathematics | S |
| Davis, Jimena | F | Clemson U | Mathematical Sciences | S |
| Douglas, Christian | M | U of North Carolina-Asheville | | S |
| Lanctot, Matthew | M | U of North Carolina-Greensboro | | S |
| Morrison, Christina | F | Clemson U | | S |
| Nutter, Amelia | F | U of North Carolina-Asheville | | S |
| Parker, Sonia | F | U of North Carolina-Wilmington | Mathematics | S |
| Pham, AnhTu | F | Murray State U | | S |
| Schoen, Rob | M | U of North Carolina-Asheville | | S |
| Sergent, Kimberly | F | U of North Carolina-Asheville | | S |
| Stroud, Lara | F | Meredith College | | S |
| Williams, Daphne | F | Clark Atlanta U | | S |
| Ziemiecki, Ryan | M | U of North Carolina-Wilmington | Statistics | S |

**Education & Outreach Program**
**SAMSI/CRSC Interdisciplinary Workshop for Undergraduates**
Campus of North Carolina State University
*Workshop Participants*
June 9-13, 2003

| Name | Gender | Affiliation | Major/Department | Status |
|---|---|---|---|---|
| Billing, Emily | F | U of North Carolina-Wilmington | Mathematics and Physics | S |
| Bakewell, Edward | M | U of North Carolina-Wilmington | Mathematics & Statistics | S |
| Charpentier, Heidi | F | U of North Carolina-Asheville | Applied Mathematics | S |
| Davis, Jimena | F | Clemson U | Mathematical Sciences | S |
| Douglas, Christian | M | U of North Carolina-Asheville | | S |
| Lanctot, Matthew | M | U of North Carolina-Greensboro | | S |
| Morrison, Christina | F | Clemson U | | S |
| Nutter, Amelia | F | U of North Carolina-Asheville | | S |
| Parker, Sonia | F | U of North Carolina-Wilmington | Mathematics | S |
| Pham, AnhTu | F | Murray State U | | S |
| Schoen, Rob | M | U of North Carolina-Asheville | | S |
| Sergent, Kimberly | F | U of North Carolina-Asheville | | S |
| Stroud, Lara | F | Meredith College | | S |
| Williams, Daphne | F | Clark Atlanta U | | S |
| Ziemiecki, Ryan | M | U of North Carolina-Wilmington | Statistics | S |

**Stochastic Computation Program**
**Closing Workshop**
*Supported Workshop Participants*
Radisson Governors Inn
June 26-28, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Besag, Julian | M | U of Washington | Statistics | FP |
| Garcia, Luis David | M | Virginia Tech U | Mathematics | NRG |
| Giudici, Paolo | M | U of Pavia | Statistics | FP |
| Goldman, Elena | F | Pace U | Business | NRG |
| Haran, Murali | M | U of Minnesota | Statistics | NRG |
| Massam, Helene | F | York U | Mathematics & Statistics | FP |
| Pasarica, Cristian | M | Columbia U | Mathematics | NRG |
| Peruggia, Mario | M | Ohio State U | Statistics | FP |
| Prendergast, Ivy | F | U of Iowa | Mathematics | NRG |
| Slavkovic, Aleksandra | F | Carnegie Mellon U | Statistics | NRG |
| Soberanis, Policarpio | M | U of Iowa | Mathematics | NRG |
| Sullivant, Seth | M | U of California-Berkeley | Mathematics | NRG |
| Yoshida, Ruriko | F | U of California-Davis | Mathematics | NRG |

**Stochastic Computation Program**
**Closing Workshop**
*Workshop Participants*
Radisson Governors Inn
June 26-28, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Bayarri, M.J. | F | U of Valencia | Statistics | FP |
| Berger, Jim | M | Duke U & SAMSI | Statistics | FP |
| Besag, Julian | M | U of Washington | Statistics | FP |
| Carter, Christopher | M | Duke U & SAMSI | Statistics | FP |
| Carvalho, Carlos | M | Duke U | Statistics | NRG |
| Chen, Yuguo | M | Duke U | Statistics | FP |
| Clyde, Merlise | F | Duke U | Statistics | FP |
| Dinwoodie, Ian | M | Duke U & SAMSI | Statistics | FP |
| Dobra, Adrian | M | Duke U & SAMSI | Statistics | NRG |
| Fouque, Jean Pierre | M | North Carolina State U | Mathematics | FP |
| Garcia, Luis David | M | Virginia Tech U | Mathematics | NRG |
| Ghosh, Sujit | M | North Carolina State U | Statistics | FP |
| Giudici, Paolo | M | U of Pavia | Statistics | FP |
| Goldman, Elena | F | Pace U | Business | NRG |
| Han, Sean | M | North Carolina State U | Mathematics | NRG |
| Hans, Chris | M | Duke U | Statistics | NRG |
| Haran, Murali | M | U of Minnesota | Statistics | NRG |

| | | | | |
|---|---|---|---|---|
| Hartemink, Alex | M | Duke U | Computer Science | FP |
| Huber, Mark | M | Duke U | Mathematics | FP |
| Ibrahim, Joe | M | U of North Carolina | Biostatistics | FP |
| Ji, Chuanshu | M | U of North Carolina | Statistics | FP |
| Jones, Beatrix | F | SAMSI | Statistics | NRG |
| Lee, Beom | M | U of North Carolina | Statistics | NRG |
| Liang, Feng | F | Duke U | Statistics | FP |
| Massam, Helene | F | York U | Mathematics & Statistics | FP |
| McCulloch, Rob | M | U of Chicago | Econometrics & Statistics | FP |
| Molina, German | M | Duke U | Statistics | NRG |
| Pang, Tao | M | North Carolina State U | Mathematics | NRG |
| Pasarica, Cristian | M | Columbia U | Mathematics | NRG |
| Paulo, Rui | M | SAMSI & NISS | Statistics | NRG |
| Peruggia, Mario | M | Ohio State U | Statistics | FP |
| Pittman, Jennifer | F | Duke U | Statistics | FP |
| Prendergast, Ivy | F | U of Iowa | Mathematics | NRG |
| Sahu, Sujit | M | U of Southampton | Mathematics | FP |
| Sanchez, Emmanuel | M | North Carolina State U | Mathematics | NRG |
| Slavkovic, Aleksandra | F | Carnegie Mellon U | Statistics | NRG |
| Soberanis, Policarpio | M | U of Iowa | Mathematics | NRG |
| Sullivant, Seth | M | U of California-Berkeley | Mathematics | NRG |

| | | | | |
|---|---|---|---|---|
| West, Mike | M | Duke U | Statistics | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |
| Wu, Yichao | M | U of North Carolina | Mathematics | NRG |
| Yoshida, Ruriko | F | U of California-Davis | Mathematics | NRG |
| Zhang, Hao (Helen) | F | North Carolina State U | Statistics | FP |

**Education & Outreach Program**
**SAMSI/CRSC Industrial Mathematical & Statistical Modeling**
**Workshop for Graduates**
Campus of North Carolina State University
*Supported Workshop Participants*
July 21-29, 2003

| **Name** | **Gender** | **Affiliation** | **Major/Department** | **Status** |
|---|---|---|---|---|
| Ananthasayanam, Balajee | M | Clemson U | | S |
| Bakewell, Edward | M | U of North Carolina-Wilmington | Mathematics & Statistics | S |
| Chen, Pengwen | M | U of Florida | | S |
| Edmonds, Bartlett | M | Virginia Commonwealth U | | S |
| Ghosh, Krishnendu | M | U of Wisconsin-Milwaukee | | S |
| Huang, Yujun | M | Georgia Institute of Technology | | S |
| Jackson, Billy | M | U of Georgia | | S |
| Jiang, Dongming | M | U of Cincinnati | | S |
| Kala, Sirish | F | Mississippi State U | | S |
| Lapin, Serguei | M | U of Houston | | S |
| Lavrik, Ilya | M | Georgia Institute of Technology | | S |

| | | | | |
|---|---|---|---|---|
| Lawler, Stacey | F | Murray State U | | S |
| Malaugh, James | M | U of Alabama-Birmingham | | S |
| Minh, Ha Quang | M | Brown U | | S |
| Nfodjo, David | M | U of Louisville | | S |
| Rus, George | M | Western Illinois U | | S |
| Sabuwala, Adnan | M | U of Florida | | S |
| Samyono, Widodo | M | Old Dominion U | | S |
| Schoen, Rob | M | U of North Carolina-Asheville | | S |
| Thatcher, Aaron | M | U of North Carolina-Wilmington | | S |
| Twagilimana, Joseph | M | U of Louisville | | S |
| White, Gentry | M | U of Missouri-Columbia | | S |
| Xu, Xiangrong | M | Florida State U | | S |
| Zeng, Bo | | Purdue U | | S |
| Zhou, Yingchun | | Boston U | | S |

**Education & Outreach Program**
**SAMSI/CRSC Industrial Mathematical & Statistical Modeling**
**Workshop for Graduates**
Campus of North Carolina State University
*Workshop Participants*
July 21-29, 2003

| **Name** | **Gender** | **Affiliation** | **Major/Department** | **Status** |
|---|---|---|---|---|
| Ananthasayanam, Balajee | M | Clemson U | | S |
| Bakewell, Edward | M | U of North Carolina-Wilmington | Mathematics & Statistics | S |

| | | | | |
|---|---|---|---|---|
| Banks, H.T. | M | North Carolina State U | CRSC | F |
| Bao, Yujuan | F | Duke U | | S |
| Barton, Hugh | M | Environmental Protection Agency | | P |
| Bergeron, Deana | F | Louisiana State U | | S |
| Buche, Robert | M | North Carolina State U | Mathematics | F |
| Chan, Edna | F | North Carolina State U | | S |
| Chen, Pengwen | M | U of Florida | | S |
| Choe, Dong-Kyoung | F | North Carolina State U | | S |
| Dausch, David | M | MCNC Research & Development Institute | | P |
| Davis, Jimena | F | Clemson U | Mathematical Sciences | S |
| Edmonds, Bartlett | M | Virginia Commonwealth U | | S |
| Ernstberger, Jon | M | Murray State U | | S |
| Fricks, John | M | U of North Carolina | | S |
| Ghosh, Krishnendu | M | U of Wisconsin-Milwaukee | | S |
| Gilbert, James | M | U of South Florida | | S |
| Goodwin, Scott | M | MCNC Research & Development Institute | | P |
| Gremaud, Pierre | M | North Carolina State U | Mathematics | F |
| Grove, Sarah | F | Youngstown State U | | S |
| Haider, Mansoor | M | North Carolina State U | Mathematics | F |
| Huang, Yujun | M | Georgia Institute of Technology | | S |
| Hunter, William | M | Fed Reserve Bank of Chicago | | P |

| | | | | |
|---|---|---|---|---|
| Ito, Kazufumi | M | North Carolina State U | Mathematics | F |
| Jackson, Billy | M | U of Georgia | | S |
| Jalali, Mohmmadreza | M | U of North Carolina-Charlotte | | S |
| Jiang, Dongming | M | U of Cincinnati | | S |
| Kala, Sirish | F | Mississippi State U | | S |
| Karr, Alan | M | National Institute of Statistical Sciences | | P |
| Lapin, Serguei | M | U of Houston | | S |
| Lavrik, Ilya | M | Georgia Institute of Technology | | S |
| Lawler, Stacey | F | Murray State U | | S |
| Li, Zhilin | M | North Carolina State U | Mathematics | F |
| Malaugh, James | M | U of Alabama-Birmingham | | S |
| Maldague, Pierre | M | Jet Propulsion Laboratory | | P |
| Minh, Ha Quang | M | Brown U | | S |
| Nfodjo, David | M | U of Louisville | | S |
| Pang, Tao | M | North Carolina State U | Mathematics | F |
| Royal, Tony | M | Jenike & Johanson, Inc | | P |
| Rus, George | M | Western Illinois U | | S |
| Sabuwala, Adnan | M | U of Florida | | S |
| Samyono, Widodo | M | Old Dominion U | | S |
| Schoen, Rob | M | U of North Carolina-Asheville | | S |
| Setzer, Woodrow | M | Environmental Protection Agency | | P |

| | | | | |
|---|---|---|---|---|
| Smith, Ralph | M | North Carolina State U | CRSC | F |
| Stroud, Lara | F | North Carolina State U | | S |
| Thatcher, Aaron | M | U of North Carolina-Wilmington | | S |
| Tran, Hien | M | North Carolina State U | Mathematics | F |
| Twagilimana, Joseph | M | U of Louisville | | S |
| White, Gentry | M | U of Missouri-Columbia | | S |
| Wong, Andrea | F | U of North Carolina-Greensboro | | S |
| Xu, Xiangrong | M | Florida State U | | S |
| Zeng, Bo | | Purdue U | | S |
| Zhou, Yingchun | | Boston U | | S |

**Data Mining and Machine Learning Program**
**Opening Workshop**
Radisson Hotel Research Triangle Park
*Supported Workshop Participants*
September 6-10, 2003

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Andries, Erik | M | U of New Mexico | Mathematics | NRG |
| Breiman, Leo | M | U of California-Berkeley | Statistics | FP |
| Chakraborty, Sounak | M | U of Florida | Statistics | NRG |
| Cook, Di | F | Iowa State U | Statistics | FP |
| DuMouchel, William | M | AT&T | | FP |
| Ghosh, Malay | M | U of Florida | Statistics | FP |

| | | | | |
|---|---|---|---|---|
| Huang, Xiaohong | F | U of Minnesota | Biostatistics | NRG |
| Jordan, Michael | M | U of California-Berkeley | Statistics | FP |
| Joshi, Saket | M | Oregon State U | Computer Science | NRG |
| Kafadar, Karen | F | U of Colorado-Denver | Mathematics | FP |
| Li, Jia | F | Pennsylvania State U | Statistics | NRG |
| Liu, Ray | M | Cornell U | | NRG |
| Madigan, David | M | Rutgers U | Statistics | FP |
| Markatou, Marianthi | F | Columbia U | Biostatistics | FP |
| Mehta, Tapan | M | U of Alabama-Birmingham | Electrical & Computer Engineering | NRG |
| Meloche, Jean | M | Avaya Labs Research | Data Analysis | FP |
| Nair, Vijay | M | U of Michigan | Statistics | FP |
| Palomo, Jesus | M | Rey Juan Carlos U | | NRG |
| Pan, Wei | | U of Minnesota | Biostatistics | |
| Quang, Minh Ha | M | Brown U | Mathematics | NRG |
| Schneider, Jeff | M | Carnegie Mellon U | Robotics Institute | FP |
| Simmons, Susan | F | U of North Carolina-Wilmington | Mathematics and Statistics | NRG |
| Sun, Jaiyang | F | Case Western Reserve U | Statistics | FP |
| Udoh, Emmanuel | M | Indiana U & Purdue U | Computer Science | FP |
| van Horebeek, Johan | M | U of Waterloo and CIMAT | Statistics & Actuarial Sciences | NRG |
| Westfall, Peter | M | Texas Tech U | | FP |
| Wolfe, Patrick | M | U of Cambridge | Engineering | NRG |

**Data Mining and Machine Learning Program**
**Opening Workshop**
Radisson Hotel Research Triangle Park
*Workshop Participants*
September 6-10, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Ahn, Jeongyoun | F | U of North Carolina | Statistics | NRG |
| Andries, Erik | M | U of New Mexico | Mathematics | NRG |
| Baek, Jong-ho | M | U of California-Los Angeles | Statistics | NRG |
| Banks, David | M | Duke U | Statistics | FP |
| Bayarri, M.J. | F | U of Valencia | Statistics & Operations Research | FP |
| Beecher, Chris | M | Metabolon | | FP |
| Berger, Jim | M | SAMSI | | FP |
| Boyer, Joe | M | North Carolina State U | Statistics | NRG |
| Breiman, Leo | M | U of California-Berkeley | Statistics | FP |
| Brooks, Atina | F | North Carolina State U | Statistics | NRG |
| Chakraborty, Sounak | M | U of Florida | Statistics | NRG |
| Chen, Yuguo | M | Duke U | Statistics | FP |
| Chu, Jen-hwa | M | Duke U | Statistics | NRG |
| Clarke, Bertrand | M | Univesity of British Columbia | Statistics | FP |
| Cook, Di | F | Iowa State U | Statistics | FP |
| Davis, Jerry | M | North Carolina State U | Marine, Earth and Atmospheric Sciences | FP |
| Davis, Jimena | F | North Carolina State U | Mathematics | NRG |

| | | | | |
|---|---|---|---|---|
| DuMouchel, William | M | AT&T | | FP |
| Eltinge, John | M | Bureau of Labor Statistics | Survey Methods Research | FP |
| Ernstberger, Jon | M | North Carolina State U | Mathematics | NRG |
| Feng, Jun | M | NISS | | NRG |
| Feng, Sheng | M | North Carolina State U | Statistics | NRG |
| Fokoue, Ernest | M | SAMSI | | NRG |
| Genton, Marc | M | North Carolina State U | Statistics | FP |
| Ghosal, Subhashis | M | North Carolina State U | Statistics | FP |
| Ghosh, Malay | M | U of Florida | Statistics | FP |
| Godden, Kurt | M | GM Research & Development | Manufacturing Systems Research | FP |
| Godtliebsen, Fred | M | U of Tromso & U of North Carolina | Statistics | FP |
| Grove, Sarah | F | North Carolina State U | Mathematics | NRG |
| Hans, Chris | M | Duke U | Statistics | NRG |
| Haran, Murali | M | NISS | | NRG |
| Hartemink, Alexander | M | Duke U | Computer Science | FP |
| Hawala, Sam | M | US Census Bureau | Statistical Research | FP |
| House, Leanna | F | Duke U | Statistics | NRG |
| Huang, Xiaohong | F | U of Minnesota | Biostatistics | NRG |
| Hughes-Oliver, Jacqueline | F | North Carolina State U | Statistics | FP |
| Hurtado, Gerardo | M | SAS Institute | Analytical Solutions | FP |
| Johnson, Brent | M | U of North Carolina | Biostatistics | FP |

| | | | | |
|---|---|---|---|---|
| Jordan, Michael | M | U of California-Berkeley | Statistics | FP |
| Joshi, Saket | M | Oregon State U | Computer Science | NRG |
| Ju, Wen-Hua | | Avaya Labs Research | | FP |
| Kafadar, Karen | F | U of Colorado, Denver | Mathematics | FP |
| Karr, Alan | M | NISS | | FP |
| Kettenring, Jon | M | Telcordia Technologies | Information Analysis and Services Research | FP |
| Kushler, Robert | M | Oakland U | Mathematics and Statistics | FP |
| LaCroix, Karol | F | GlaxoSmithKline Inc. | Safety | FP |
| Lee, Taiyeong | M | SAS Institute | | FP |
| Li, Jia | F | Pennsylvania State U | Statistics | NRG |
| Liang, Feng | F | Duke U | Statistics | FP |
| Liggett, Walter | M | Natl Inst of Standards & Technology | Statistical Engineering | FP |
| Lin, Danyu | F | U of North Carolina | Biostatistics | NRG |
| Lin, Xiaodong | M | SAMSI | | NRG |
| Liu, Fei | F | Duke U | Statistics | NRG |
| Liu, Jack | M | NISS | | NRG |
| Liu, Peng | M | North Carolina State U | Statistics | NRG |
| Liu, Ray | M | Cornell U | | NRG |
| Luedi, Philippe | M | Duke U | Bioinformatics | NRG |
| Madigan, David | M | Rutgers U | Statistics | FP |
| Markatou, Marianthi | F | Columbia U | Biostatistics | FP |

| | | | | |
|---|---|---|---|---|
| Marron, Steve | M | SAMSI | | FP |
| McCulloch, Robert | M | U of Chicago | Graduate School of Business | FP |
| Mehta, Tapan | M | U of Alabama, Birmingham | Electrical & Computer Engineering | NRG |
| Meloche, Jean | M | Avaya Labs Research | Data Analysis | FP |
| Michailidis, George | M | U of Michigan | Statistics | FP |
| Montgomery, Jr., Paul | M | GM Research & Development | Manufacturing Systems Research Lab | FP |
| Nair, Vijay | M | U of Michigan | Statistics | FP |
| Nobel, Andrew | M | U of North Carolina | Statistics | FP |
| Noe, Doug | M | U of Illinois, Urbana-Champaign | Statistics | NRG |
| Palomo, Jesus | M | Rey Juan Carlos U | | NRG |
| Pan, Wei | | U of Minnesota | Biostatistics | |
| Park, Cheolwoo | M | SAMSI | Statistics | NRG |
| Park, Juhyun | F | U of North Carolina | Statistics | NRG |
| Park, Soyoun | F | U of North Carolina | Statistics | NRG |
| Peddada, Shyamal | M | NIEHS | Biostatistics | FP |
| Quang, Minh Ha | M | Brown U | Mathematics | NRG |
| Randolph, Timothy | M | U of Washington | Biostatistics | FP |
| Rangel, Laureano | M | North Carolina State U | Statistics | NRG |
| Remlinger, Katja | F | North Carolina State U | Statistics | NRG |
| Rodriguez, Robert | M | SAS Institute | Analytical Solutions | FP |
| Sanil, Ashish | M | NISS | | FP |

| | | | | |
|---|---|---|---|---|
| Schneider, Jeff | M | Carnegie Mellon U | Robotics Institute | FP |
| Simmons, Susan | F | U of North Carolina-Wilmington | Mathematics and Statistics | NRG |
| Simpson, Doug | M | U of Illinois-Urbana-Champaign | Statistics | FP |
| Srivastava, Vaibhav | M | North Carolina State U | Electrical & Computer Engineering | NRG |
| Sun, Dongchu | M | U of Missouri | Statistics | FP |
| Sun, Jaiyang | F | Case Western Reserve U | Statistics | FP |
| Taqqu, Murad | M | SAMSI | | FP |
| Thibaudeau, Yves | M | Statistical Research Census Bureau | Statistical Research Division | FP |
| Truong, Young | M | U of North Carolina | Biostatistics | FP |
| Udoh, Emmanuel | M | Indiana U & Purdue U | Computer Science | FP |
| Unnikrishnan, K.P. | M | GM Research & Development | | FP |
| van Horebeek, Johan | M | U of Waterloo and CIMAT | Statistics & Actuarial Sciences | NRG |
| van Rhee, Michiel | M | ICAGEN Inc. | Discovery Chemistry | NRG |
| Vance, Eric | M | Duke U | Statistics | NRG |
| Wang, Wei | F | U of North Carolina | Computer Science | FP |
| Wang, Xiaohui | | U of North Carolina | Statistics | NRG |
| Westfall, Peter | M | Texas Tech U | | FP |
| Wolfe, Patrick | M | U of Cambridge | Engineering | NRG |
| Wu, Yujun | | North Carolina State U | Statistics | NRG |
| Young, Stan | M | NISS | | FP |
| Yuen, Nancy | F | GlaxoSmithKline Inc. | Clinical Safety | FP |

| | | | | |
|---|---|---|---|---|
| Zhang, Hao | F | North Carolina State U | Statistics | NRG |
| Zhang, Ke | M | North Carolina State U | Statistics | NRG |
| Zhu, Kaiding | M | U of CA, Los Angeles | Statistics | NRG |
| Zhu, Zhengyuan | M | U of North Carolina | Statistics | FP |

**Network Modeling for the Internet Program**
**Workshops on Internet Tomography and Sensor Networks**
Radisson Hotel Research Triangle Park
*Supported Workshop Participants*
October 12-15, 2003

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Abry, Patrice | M | CNRS | Physics Laboratory | FP |
| Baraniuk, Richard | M | Rice U | Electrical & Computer Engineering | FP |
| Broido, Andre | M | CAIDA-SDSC-UCSD | | FP |
| Castro, Rui | M | Rice U | Electrical & Computer Engineering | NRG |
| Cetin, Mujdat | M | MIT | Lab for Information & Decision Systems | FP |
| Chua, David | M | Boston U | Mathematics | NRG |
| Coates, Mark | M | McGill U | Electrical & Computer Engineering | FP |
| Crovella, Mark | M | Boston U | Computer Science | FP |
| Cruz, Rene | M | U of California-San Diego | Electrical & Computer Engineering | FP |
| Denby, Lorraine | F | Avaya | Labs Research | FP |
| Duffield, Nicholas | M | AT&T Research | | FP |
| Estrin, Deborah | F | U of California-Los Angeles & CENS | Computer Science | FP |

| | | | | |
|---|---|---|---|---|
| Fisher, John | M | MIT | Electrical Engineering & Computer Sciences | FP |
| Giles, Kendall | M | Johns Hopkins U | Computer Science | NRG |
| Grunbaum, Alberto | M | U of California-Berkeley | Mathematics | FP |
| Hero, Alfred | M | U of Michigan | Electrical Engineering & Computer Sciences | FP |
| Jun, Mikyoung | F | U of Chicago | Statistics | NRG |
| Kolaczyk, Eric | M | Boston U | Mathematics & Statistics | FP |
| Lakhina, Anukool | M | Boston U | Computer Science | NRG |
| Landwehr, Jim | M | Avaya | Data Analysis Research | FP |
| Lawrence, Earl | M | U of Michigan | Statistics | NRG |
| Lee, Thomas | M | Colorado State U | Statistics | FP |
| Lehoczky, John | M | Carnegie Mellon U | Statistics | FP |
| Liang, Gang | M | U of California-Berkeley | Statistics | NRG |
| Liu, Benyuan | M | City College of New York | Computer Science | FP |
| Makowski, Armand | M | U of Maryland | Electrical & Computer Engineering | FP |
| Meloche, Jean | M | Avaya | Data Analysis Research | FP |
| Moura, Jose | M | Carnegie Mellon U | Electrical & Computer Engineering | FP |
| Nair, Vijay | M | U of Michigan | Statistics | FP |
| Nucci, Antonio | M | Sprintlabs | IP Group | NRG |
| Papadopoulos, Christos | M | U of Southern CA & Information Sci Inst | Computer Science | FP |
| Papagiannaki, Konstantina | F | Sprint ATL | | FP |
| Rahimi, Mohammad | M | U of California-Los Angeles | Center for Embedded Networked Sensing | NRG |

| | | | | |
|---|---|---|---|---|
| Ravishanker, Nalini | F | U of Connecticut | Statistics | FP |
| Roughan, Matthew | M | AT&T Research | | FP |
| Sommers, Joel | M | U of Wisconsin | Computer Science | NRG |
| Srikant, R | M | U of Illinois | | FP |
| Srivastava, Mani | M | U of California-Los Angeles | Electrical Engineering | FP |
| Sun, Jiayang | F | Case Western Reserve U | Statistics | FP |
| Tsang, Yolanda | F | Rice U | Electrical & Computer Engineering | NRG |
| Wei, Wei | M | U of Massachusetts | Computer Science | NRG |
| Weng, Jing | | U of Massachusetts-Amherst | Computer Science | NRG |
| Xia, Cathy | F | IBM Research | Distributed Computing | FP |
| Xi, Bowei | F | U of Michigan-Ann Arbor | Statistics | NRG |
| Yao, Kung | M | U of California-Los Angeles | Electrical Engineering | FP |
| Yegneswaran, Vinod | M | U of Wisconsin-Madison | Computer Science | NRG |
| Yu, Bin | F | U of California-Berkeley | Statistics | FP |
| Zhao, Feng | M | Palo Alto Research Center | Systems & Practices Laboratory | FP |
| Zhao, Linda | F | U of Pennsylvania | Statistics | FP |

**Network Modeling for the Internet Program**
**Workshops on Internet Tomography and Sensor Networks**
Radisson Hotel Research Triangle Park
*Workshop Participants*
October 12-15, 2003

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Abry, Patrice | M | CNRS | Physics Laboratory | FP |
| Baraniuk, Richard | M | Rice U | Electrical & Computer Engineering | FP |
| Bragg, Arnold | M | MCNC RDI | Advanced Network Research | FP |
| Broido, Andre | M | CAIDA-SDSC-UCSD | | FP |
| Buche, Robert | M | North Carolina State U | Mathematics | FP |
| Castro, Rui | M | Rice U | Electrical & Computer Engineering | NRG |
| Cetin, Mujdat | M | MIT | Lab for Information & Decision Systems | FP |
| Chek, John | M | U of North Carolina | Computer Science | NRG |
| Chinchilla, Francisco | M | U of North Carolina | Computer Science | NRG |
| Chua, David | M | Boston U | Mathematics | NRG |
| Coates, Mark | M | McGill U | Electrical & Computer Engineering | FP |
| Crovella, Mark | M | Boston U | Computer Science | FP |
| Cruz, Rene | M | U of California-San Diego | Electrical & Computer Engineering | FP |
| Denby, Lorraine | F | Avaya | Labs Research | FP |
| Devetsikiotis, Michael | M | North Carolina State U | Electrical & Computer Engineering | FP |
| Dewan, Prasun | M | U of North Carolina | Computer Science | FP |
| Dinwoodie, Ian | M | Duke U | Statistics | FP |

| | | | | |
|---|---|---|---|---|
| Duffield, Nicholas | M | AT&T Research | | FP |
| Estrin, Deborah | F | U of California-Los Angeles & CENS | Computer Science | FP |
| Fisher, John | M | MIT | Electrical Engineering & Computer Sciences | FP |
| Ghosh, Arka | M | U of North Carolina | | NRG |
| Giles, Kendall | M | Johns Hopkins U | Computer Science | NRG |
| Godtliebsen, Fred | M | U of North Carolina | Statistics & Operations Research | FP |
| Grunbaum, Alberto | M | U of California-Berkeley | Mathematics | FP |
| Hansen, Mark | M | U of California-Los Angeles | Statistics | |
| Harfoush, Khaled | M | North Carolina State U | Computer Science | FP |
| Hero, Alfred | M | U of Michigan | Electrical Engineering & Computer Sciences | FP |
| Izem, Rima | F | U of North Carolina | Statistics & Operations Research | NRG |
| Jeffay, Kevin | M | U of North Carolina | Computer Science | FP |
| Jun, Mikyoung | F | U of Chicago | Statistics | NRG |
| Karr, Alan | M | NISS | | FP |
| Kaur, Jasleen | F | U of North Carolina | Computer Science | FP |
| Kolaczyk, Eric | M | Boston U | Mathematics & Statistics | FP |
| Lakhina, Anukool | M | Boston U | Computer Science | NRG |
| Landwehr, Jim | M | Avaya | Data Analysis Research | FP |
| Lawrence, Earl | M | U of Michigan | Statistics | NRG |
| Lee, Thomas | M | Colorado State U | Statistics | FP |
| Lehoczky, John | M | Carnegie Mellon U | Statistics | FP |

| | | | | |
|---|---|---|---|---|
| Levy, Josh | M | U of North Carolina | Computer Science | NRG |
| Liang, Gang | M | U of CA-Berkeley | Statistics | NRG |
| Liu, Benyuan | M | City College of New York | Computer Science | FP |
| Liu, Zhen | M | IBM Research | T.J. Watson Research Center | FP |
| Makowski, Armand | M | U of Maryland | Electrical & Computer Engineering | FP |
| Marchette, David | M | United States Navy | Naval Surface Warfare Center | FP |
| Maulik, Krishanu | M | EURANDOM | | FP |
| Meloche, Jean | M | Avaya | Data Analysis Research | FP |
| Michailidis, George | M | U of Michigan | Statistics | FP |
| Moura, Jose | M | Carnegie Mellon U | Electrical & Computer Engineering | FP |
| Nair, Vijay | M | U of Michigan | Statistics | FP |
| Nilsson, Arne | M | North Carolina State U | Electrical & Computer Engineering | FP |
| Nobel, Andrew | M | U of North Carolina | Statistics & Operations Research | FP |
| Nowak, Rob | M | U of Wisconsin | Electrical & Computer Engineering | FP |
| Nucci, Antonio | M | Sprintlabs | IP Group | NRG |
| Papadopouli, Maria | F | U of North Carolina | Computer Science | FP |
| Papadopoulos, Christos | M | U of Southern CA & Information Sci Inst | Computer Science | FP |
| Papagiannaki, Konstantina | F | Sprint ATL | | FP |
| Park, Cheolwoo | M | SAMSI | | NRG |
| Park, Juhyun | F | U of North Carolina | Statistics & Operations Research | NRG |
| Park, Sang Joon | | North Carolina State U | Electrical & Computer Engineering | NRG |

| | | | | |
|---|---|---|---|---|
| Pipiras, Vladas | M | U of North Carolina | Statistics & Operations Research | FP |
| Rahimi, Mohammad | M | U of California-Los Angeles | Center for Embedded Networked Sensing | NRG |
| Ramchandran, Kannan | M | U of California-Berkeley | Electrical Engineering & Computer Sciences | FP |
| Ravishanker, Nalini | F | U of Connecticut | Statistics | FP |
| Rewaskar, Sushant | M | U of North Carolina | Computer Science | NRG |
| Rolls, David | M | SAMSI & U of North Carolina-Wilmington | Mathematics & Statistics | FP |
| Roughan, Matthew | M | AT&T Research | | FP |
| Sabbineni, Harshavardhan | M | Duke U | Electical Engineering | NRG |
| Shen, Haipeng | M | U of North Carolina | Statistics & Operations Research | NRG |
| Shriram, Alok | M | U of North Carolina | Computer Science | NRG |
| Smith, Don | M | U of North Carolina | Computer Science | FP |
| Sommers, Joel | M | U of Wisconsin | Computer Science | NRG |
| Srikant, R | M | U of Illinois | | FP |
| Srivastava, Mani | M | U of California-Los Angeles | Electrical Engineering | FP |
| Stoev, Stilian | M | Boston U | Mathematics & Statistics | NRG |
| Sun, Jiayang | F | Case Western Reserve U | Statistics | FP |
| Taqqu, Murad | M | SAMSI & Boston U | CAS Mathematics & Statistics | FP |
| Towsley, Don | M | U of Massachusetts | Computer Science | FP |
| Trussell, Joel | M | North Carolina State U | Electrical & Computer Engineering | FP |
| Tsang, Yolanda | F | Rice U | Electrical & Computer Engineering | NRG |
| Vernon, Frank | M | U of California-San Diego | Institute of Geophysics & Planetary Physics | FP |

| | | | | |
|---|---|---|---|---|
| Wang, Cliff | M | Army Research Office | Computing & Information Sci Division | FP |
| Wei, Wei | M | U of Massachusetts | Computer Science | NRG |
| Weng, Jing | | U of Massachusetts-Amherst | Computer Science | NRG |
| Willinger, Walter | M | AT&T Research | | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |
| Xia, Cathy | F | IBM Research | Distributed Computing | FP |
| Xi, Bowei | F | U of Michigan-Ann Arbor | Statistics | NRG |
| Yao, Kung | M | U of California-Los Angeles | Electrical Engineering | FP |
| Yegneswaran, Vinod | M | U of Wisconsin-Madison | Computer Science | NRG |
| Yu, Bin | F | U of California-Berkeley | Statistics | FP |
| Zhao, Feng | M | Palo Alto Research Center | Systems & Practices Laboratory | FP |
| Zhao, Linda | F | U of Pennsylvania | Statistics | FP |
| Zhu, Zhengyuan | M | U of North Carolina | Statistics & Operations Research | FP |
| Zou, Yi | M | Duke U | Electrical & Computer Engineering | NRG |

**Network Modeling for the Internet Program**
**Workshops on Congestion Control and Heavy Traffic Modeling**
Radisson Hotel Research Triangle Park
*Supported Workshop Participants*
October 31-November 1, 2003

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Cao, Jin | F | Bell Labs and Lucent Technologies | Statistics and Data Mining Research | FP |
| Dai, Jim | M | Georgia Institute of Technology | Industrial and Systems Engineering | FP |

| | | | | |
|---|---|---|---|---|
| D'Auria, Bernardo | M | Cornell U | Operations Research & Industrial Engineering | NRG |
| Gamundi, Emily | F | Tulane U | Mathematics | NRG |
| Glynn, Peter | M | Stanford U | Management Science and Engineering | FP |
| Hannig, Jan | M | Colorado State U | Statistics | |
| Harmantzis, Fotios | M | Stevens Institute of Technology | | FP |
| Karagiannis, Thomas | M | CAIDA and U of California-Riverside | Computer Science | FP |
| Lehoczky, John | M | Carnegie Mellon U | Statistics | FP |
| Makowski, Armand | M | U of Maryland | Electrical and Computer Engineering | FP |
| Misra, Vishal | M | Columbia U | Computer Science | FP |
| Samorodnitsky, Gennady | M | Cornell U | Operations Research & Industrial Engineering | FP |
| Shah, Khushboo | F | U of Southern California | | NRG |
| Zhang, Li | | IBM | T.J. Watson Research Center | FP |

**Network Modeling for the Internet Program**
**Workshops on Congestion Control and Heavy Traffic Modeling**
Radisson Hotel Research Triangle Park
*Workshop Participants*
October 31-November 1, 2003

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Ahn, Jeongyoun | F | U of North Carolina | Statistics and Operations Research | NRG |
| Akin, Ozdemir | | North Carolina State U | Electrical and Computer Engineering | NRG |
| Baccelli, Francois | M | ENS | | FP |
| Bohacek, Stephen | M | U of Delaware | Electrical and Computer Engineering | FP |

| | | | | |
|---|---|---|---|---|
| Buche, Robert | M | North Carolina State U | Mathematics | FP |
| Budhiraja, Amarjit | M | U of North Carolina | Statistics and Operations Research | FP |
| Cao, Jin | F | Bell Labs and Lucent Technologies | Statistics and Data Mining Research | FP |
| Cleveland, William | M | Bell Labs | Statistics Research | FP |
| Dai, Jim | M | Georgia Institute of Technology | Industrial and Systems Engineering | FP |
| D'Auria, Bernardo | M | Cornell U | Operations Research & Industrial Engineering | NRG |
| Devetsikiotis, Michael | M | North Carolina State U | Electrical and Computer Engineering | FP |
| Dinwoodie, Ian | M | Duke U | Statistics | FP |
| Eun, Do Young | M | North Carolina State U | Electrical and Computer Engineering | FP |
| Gamundi, Emily | F | Tulane U | Mathematics | NRG |
| Ghosh, Arka | M | U of North Carolina & SAMSI | Statistics | NRG |
| Glynn, Peter | M | Stanford U | Management Science and Engineering | FP |
| Godtliebsen, Fred | M | U of North Carolina | Statistics | FP |
| Hannig, Jan | M | Colorado State U | Statistics | |
| Harmantzis, Fotios | M | Stevens Institute of Technology | | FP |
| Hernandez-Campos, Felix | M | U of North Carolina & SAMSI | Computer Science | NRG |
| Jeffay, Kevin | M | U of North Carolina | Computer Science | FP |
| Jennings, Otis | M | Duke U | Operations Management | FP |
| Karagiannis, Thomas | M | CAIDA and U of California-Riverside | Computer Science | FP |
| Kaur, Jasleen | F | U of North Carolina | Computer Science | FP |
| Kulkarni, Vidyadhar | M | U of North Carolina | Statistics and Operations Research | FP |

| | | | | |
|---|---|---|---|---|
| Le, Long | | U of North Carolina | Computer Science | NRG |
| Lehoczky, John | M | Carnegie Mellon U | Statistics | FP |
| Lin, Chuan | M | North Carolina State U | Operations Research | NRG |
| Liu, Liqiang | | U of North Carolina | Statistics and Operations Research | NRG |
| Lu, John | M | NIST | Statistical Engineering | FP |
| Makowski, Armand | M | U of Maryland | Electrical and Computer Engineering | FP |
| Maulik, Krishanu | M | EURANDOM | | FP |
| Michailidis, George | M | U of Michigan | Statistics | FP |
| Misra, Vishal | M | Columbia U | Computer Science | FP |
| Nalatwad, Srikant | | North Carolina State U | Electrical and Computer Engineering | NRG |
| Ninan, Bobby | M | North Carolina State U | Operations Research | NRG |
| Orguganti, Swaroop | | North Carolina State U | Electrical and Computer Engineering | NRG |
| Park, Cheolwoo | M | SAMSI | | NRG |
| Park, Juhyun | F | U of North Carolina | Statistics | NRG |
| Perros, Harry | M | North Carolina State U | Computer Science | NRG |
| Pipiras, Vladas | M | U of North Carolina | Statistics | FP |
| Ray, Surajit | M | Pennsylvania State U | Statistics | FP |
| Resnick, Sidney | M | Cornell U | Operations Research & Industrial Engineering | FP |
| Rewaskar, Sushant | M | U of North Carolina | Computer Science | NRG |
| Riedi, Rudolf | M | Rice U | Statistics and Electrical & Computer Eng | FP |
| Rolls, David | M | SAMSI & U of North Carolina-Wilmington | | NRG |

| | | | | |
|---|---|---|---|---|
| Samorodnitsky, Gennady | M | Cornell U | Operations Research & Industrial Engineering | FP |
| Shah, Khushboo | F | U of Southern California | | NRG |
| Shen, Haipeng | M | U of North Carolina | Statistics and Operations Research | NRG |
| Smith, Don | M | U of North Carolina | Computer Science | FP |
| Smith, Richard | M | U of North Carolina | Statistics and Operations Research | FP |
| Stoev, Stilian | M | Boston U | Mathematics and Statistics | NRG |
| Taqqu, Murad | M | Boston U & SAMSI | Mathematics | FP |
| Towsley, Don | M | U of Massachusetts | Computer Science | FP |
| Venkatesh, Rahul | | North Carolina State U | Mathematics | NRG |
| Wang, Xiaohui | M | U of North Carolina | Statistics | NRG |
| Williams, Ruth | F | U of California-San Diego | Mathematics | FP |
| Wolpert, Robert | M | Duke U | Statistics | FP |
| Zhang, Li | M | IBM | T.J. Watson Research Center | FP |
| Zhang, Lingsong | | U of North Carolina | Statistics and Operations Research | NRG |
| Zhu, Zhengyuan | M | U of North Carolina | Statistics | FP |
| Ziya, Serhan | | U of North Carolina | Statistics and Operations Research | FP |

**Education & Outreach Program**
**Two-Day Undergraduate Workshop on Data Mining:**
**Handling the Flood of Data**
NISS-SAMSI Building
*Supported Workshop Participants*
November 14-15, 2003

| Name | Gender | Affiliation | Major/Department | Status |
|---|---|---|---|---|
| Aye, Thida | F | Bryn Mawr College | Mathematics, Physics, and Economics | S |
| Billing, Emily | F | U of North Carolina-Wilmington | Mathematics and Physics | S |
| Blanding, Carletha | F | U of North Carolina-Wilmington | Biology | S |
| Bristol, David | M | U of North Carolina-Wilmington | Mathematics | S |
| Charpentier, Heidi | F | U of North Carolina-Asheville | Applied Mathematics | S |
| Clemons, Noah | M | Vanderbilt U | Computer Science and Mathematics | S |
| Ferguson, Angela | F | U of Arkansas-Little Rock | Mathematics | S |
| Ferraro, Michael | M | U of North Carolina-Wilmington | Mathematics | S |
| Gant, Raymond | M | Benedict College | Mathematics | S |
| Gustafson, Katie | F | Truman State U | Mathematics | S |
| Iyangar, Balaji | M | Benedict College | Mathematics & Computer Science | F |
| Holmes, Tracey | F | California State U, Chico | Mathematics | S |
| Jeffries, Robin | F | California State U, Chico | Statistics and Biology | S |
| Karlon, Kathleen Mary | F | U of North Carolina-Wilmington | Mathematics | S |
| Komolafe, Olaide | F | Benedict College | Computer Science | S |
| Lambdin, Jennifer | F | Salisbury U | Mathematics | S |
| Maynard, Jennifer | F | U of North Carolina-Asheville | Pure Mathematics | S |

| | | | | |
|---|---|---|---|---|
| Meadows, Jonathan | M | U of North Carolina-Asheville | Pure Mathematics | S |
| Miracle, Calvin | M | U of Louisville | Mathematics | S |
| Nistor, Cristina | F | Bryn Mawr College | Mathematics | S |
| Pratt, Rebecca | F | Suny Oneonta | Statistics | S |
| Sando, Julia | F | U of Rochester | Applied Mathematics | S |
| Shideed, Marwa | F | College of Charleston | Computer Science and Discrete Mathematics | S |
| Tarin, Sara | F | Bryn Mawr College | Mathematics | S |
| Teicher, Melissa | F | Bryn Mawr College | Mathematics | S |
| Ward, Courtney | F | California State U, Chico | Applied Mathematics | S |
| Winkler, David | M | College of Charleston | Computer Science | S |
| Zhan, Qian "Cindy" | F | Bryn Mawr College | Mathematics and Biology | S |

**Education & Outreach Program**
**Two-Day Undergraduate Workshop on Data Mining:**
**Handling the Flood of Data**
NISS-SAMSI Building
*Workshop Participants*
November 14-15, 2003

| **Name** | **Gender** | **Affiliation** | **Major/Department** | **Status** |
|---|---|---|---|---|
| Aye, Thida | F | Bryn Mawr College | Mathematics, Physics, and Economics | S |
| Billing, Emily | F | U of North Carolina-Wilmington | Mathematics and Physics | S |
| Blanding, Carletha | F | U of North Carolina-Wilmington | Biology | S |
| Bristol, David | M | U of North Carolina-Wilmington | Mathematics | S |
| Charpentier, Heidi | F | U of North Carolina-Asheville | Applied Mathematics | S |

| | | | | |
|---|---|---|---|---|
| Clemons, Noah | M | Vanderbilt U | Computer Science and Mathematics | S |
| Ferguson, Angela | F | U of Arkansas-Little Rock | Mathematics | S |
| Ferraro, Michael | M | U of North Carolina-Wilmington | Mathematics | S |
| Gant, Raymond | M | Benedict College | Mathematics | S |
| Gustafson, Katie | F | Truman State U | Mathematics | S |
| Iyangar, Balaji | M | Benedict College | Mathematics & Computer Science | F |
| Holmes, Tracey | F | California State U, Chico | Mathematics | S |
| Jeffries, Robin | F | California State U, Chico | Statistics and Biology | S |
| Karlon, Kathleen Mary | F | U of North Carolina-Wilmington | Mathematics | S |
| Komolafe, Olaide | F | Benedict College | Computer Science | S |
| Lambdin, Jennifer | F | Salisbury U | Mathematics | S |
| Maynard, Jennifer | F | U of North Carolina-Asheville | Pure Mathematics | S |
| Meadows, Jonathan | M | U of North Carolina-Asheville | Pure Mathematics | S |
| Miracle, Calvin | M | U of Louisville | Mathematics | S |
| Nistor, Cristina | F | Bryn Mawr College | Mathematics | S |
| Pratt, Rebecca | F | Suny Oneonta | Statistics | S |
| Sando, Julia | F | U of Rochester | Applied Mathematics | S |
| Shideed, Marwa | F | College of Charleston | Computer Science and Discrete Mathematics | S |
| Tarin, Sara | F | Bryn Mawr College | Mathematics | S |
| Teicher, Melissa | F | Bryn Mawr College | Mathematics | S |
| Townsend, Howard | M | Benedict College | Computer Science | S |

| | | | | |
|---|---|---|---|---|
| Ward, Courtney | F | California State U, Chico | Applied Mathematics | S |
| Whitney, Kofi | M | Benedict College | Computer Science | S |
| Winkler, David | M | College of Charleston | Computer Science | S |
| Zhan, Qian "Cindy" | F | Bryn Mawr College | Mathematics and Biology | S |

**Multiscale Model Development and Control Design Program**
**Opening Workshop**
Radisson Hotel Research Triangle Park
*Supported Workshop Participants*
January 17-20, 2004

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Buranathiti, Thaweepat | M | Northwestern U | Mechanical Engineering | NRG |
| Crawford, Carol Gotway | F | Centers for Disease Control | National Center for Environmental Health | FP |
| Dapino, Marcelo | M | Ohio State U | Mechanical Engineering | FP |
| Hamzi, Boumedine | M | U of California-Davis | Mathematics | FP |
| Herdic, Scott | M | Georgia Institute of Technology | Mechanical Engineering | NRG |
| Higdon, David | M | Los Alamos National Laboratory | Statistics | FP |
| Hou, Tom | M | California Institute of Technology | Applied Mathematics | FP |
| Jang, Bongsoo | M | U of North Carolina-Charlotte | Mathematics | NRG |
| Jeong, Jae Woo | M | U of North Carolina-Charlotte | Mathematics | NRG |
| Johannesson, Gardar | M | Lawrence Livermore National Lab | Systems & Decision Sciences | FP |
| King, Belinda | F | Oregon State U | Mechanical Engineering | FP |
| Kloucek, Petr | M | Rice U | CAAM | FP |

| | | | | |
|---|---|---|---|---|
| Krueger, Denise | F | Virginia Tech | ICAM | NRG |
| Ladd, Josh | M | Colorado State U | Mathematics | NRG |
| Lee, Chun Man | M | Colorado State U | Statistics | FP |
| Liu, Tieqi | | Georgia Institute of Technology | Mechanical Engineering | NRG |
| Luo, Haoxiang | M | U of California-San Diego | Mechanical & Aerospace Engineering | NRG |
| Lynch, Christopher | M | Georgia Institute of Technology | Mechanical Engineering | FP |
| Mabuchi, Hideo | M | California Institute of Technology | Physics and Control & Dynamical Systems | FP |
| Melikhov, Yevgen | M | Iowa State U | Ames Laboratory | FP |
| Morris, Kirsten | F | U of Waterloo | Applied Mathematics | FP |
| Nguyen, Hoan | F | Virginia Tech | Mathematics | NRG |
| Oates, Williams | M | Georgia Institute of Technology | Mechanical Engineering | NRG |
| Oh, JinHyoung | M | U of Michigan | Aerospace Engineering | NRG |
| Prendergast, Ivy | F | U of Iowa | Applied Mathematics | NRG |
| Reynolds, Daniel | M | Lawrence Livermore National Lab | Center for Applied Scientific Computing | FP |
| Salapaka, Murti | M | Iowa State U | Electrical & Computing Engineering | FP |
| Singler, John | M | ICAM - Virginia Tech | Mathematics | NRG |
| Soberanis, Policarpio | M | U of Iowa | Applied Mathematics | NRG |
| van Handel, Ramon | M | California Institute of Technology | Physics | NRG |
| Vannucci, Marina | F | Texas A&M U | Statistics | FP |
| Vidakovic, Brani | M | Georgia Institute of Technology | Industrial & Systems Engineering | FP |
| Wang, Haonan | M | Colorado State U | Statistics | FP |

| | | | | |
|---|---|---|---|---|
| Wang, Yazhen | M | U of Connecticut | Statistics | FP |
| Wikle, Christopher | M | U of Missouri-Columbia | Statistics | FP |
| Wright, Grady | M | U of Utah | Mathematics | FP |
| Wu, C.F. Jeff | M | Georgia Institute of Technology | Industrial & Systems Engineering | FP |
| Zietsman, Lizette | F | George Mason U | Mathematical Sciences | FP |

**Multiscale Model Development and Control Design Program**
**Opening Workshop**
Radisson Hotel Research Triangle Park
*Workshop Participants*
January 17-20, 2004

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Andjelkovic, Ivan | M | North Carolina State U | Applied Mathematics | NRG |
| Banks, H.T. | M | North Carolina State U & SAMSI | CRSC | FP |
| Begashaw, Negash | M | Benedict College | Mathematics and Computer Science | FP |
| Berger, Jim | M | SAMSI | | FP |
| Bokil, Vrushali | F | North Carolina State U | CRSC | FP |
| Borggaard, Jeff | M | Virginia Tech | Mathematics | FP |
| Brenner, David | M | North Carolina State U | Materials Science & Engineering | FP |
| Buranathiti, Thaweepat | M | Northwestern U | Mechanical Engineering | NRG |
| Cioranescu, Doina | F | CNRS-U Paris 6 | Laboratoire J. L. Lions | FP |
| Cox, Dennis | M | Rice U | Statistics | FP |
| Crawford, Carol Gotway | F | Centers for Disease Control | National Center for Environmental Health | FP |

| | | | | |
|---|---|---|---|---|
| Dapino, Marcelo | M | Ohio State U | Mechanical Engineering | FP |
| Davis, Jimena | F | North Carolina State U | Mathematics | NRG |
| Ellwein, Laura | F | North Carolina State U | Mathematics | NRG |
| Ernstberger, Jon | M | North Carolina State U | CRSC & Applied Mathematics | NRG |
| Forest, Gregory | M | U of North Carolina | Mathematics & Institute for Advanced Materials | FP |
| Fuentes, Montserrat | F | North Carolina State U | Statistics | FP |
| Ganapathysubramanian, Shankar | M | Cornell U | Mechanical Engineering | NRG |
| Gelfand, Alan | M | Duke U | Statistics | FP |
| Grove, Sarah | F | North Carolina State U | Mathematics | NRG |
| Hamzi, Boumedine | M | U of California-Davis | Mathematics | FP |
| Hatch, Andrew | M | North Carolina State U | Mathematics | NRG |
| Herdic, Scott | M | Georgia Institute of Technology | Mechanical Engineering | NRG |
| Higdon, David | M | Los Alamos National Laboratory | Statistics | FP |
| Hou, Tom | M | California Institute of Technology | Applied Mathematics | FP |
| Jang, Bongsoo | M | U of North Carolina-Charlotte | Mathematics | NRG |
| Jang, Seonhee | | North Carolina State U | Materials Science & Engineering | NRG |
| Jeong, Jae Woo | M | U of North Carolina-Charlotte | Mathematics | NRG |
| Ji, Chuanshu | M | U of North Carolina | Statistics | FP |
| Johannesson, Gardar | M | Lawrence Livermore National Lab | Systems & Decision Sciences | FP |
| Kelley, Carl T. | M | North Carolina State U | Mathematics | FP |
| Kevrekidis, Yannis | M | Princeton U | Chemical Engineering | FP |

| | | | | |
|---|---|---|---|---|
| King, Belinda | F | Oregon State U | Mechanical Engineering | FP |
| Kloucek, Petr | M | Rice U | CAAM | FP |
| Krener, Arthur | M | U of California-Davis | Mathematics | FP |
| Krueger, Denise | F | ICAM - Virginia Tech | | NRG |
| Lada, Emily | F | SAMSI | | NRG |
| Ladd, Josh | M | Colorado State U | Mathematics | NRG |
| Lee, Chun Man | M | Colorado State U | Statistics | FP |
| Lee, Joo Hee | | U of North Carolina | Mathematics | NRG |
| Lee, Myung Hee | F | U of North Carolina | Statistics | NRG |
| Leo, Donald | M | Virginia Tech | Mechanical Engineering | FP |
| Li, Yanxin | | North Carolina State U | Materials Science & Engineering | FP |
| Liu, Delong | M | NIEHS and NIH | | FP |
| Liu, Tieqi | | Georgia Institute of Technology | Mechanical Engineering | NRG |
| Lu, Hongcheng | M | Northwestern U | Mechanical Engineering | FP |
| Lucas, Joe | M | Duke U | Statistics | NRG |
| Luo, Haoxiang | M | U of California-San Diego | Mechanical & Aerospace Engineering | NRG |
| Lynch, Christopher | M | Georgia Institute of Technology | Mechanical Engineering | FP |
| Mabuchi, Hideo | M | California Institute of Technology | Physics and Control & Dynamical Systems | FP |
| Matthews, Jessica | F | North Carolina State U | Mathematics | NRG |
| Melikhov, Yevgen | M | Iowa State U | Ames Laboratory | FP |
| Morris, Kirsten | F | U of Waterloo | Applied Mathematics | FP |

| | | | | |
|---|---|---|---|---|
| Mullins, William | M | U.S. Army Research Office | | FP |
| Newell, Andrew | M | North Carolina State U | CRSC | FP |
| Nguyen, Hoan | F | Virginia Tech | Mathematics | NRG |
| Oates, Williams | M | Georgia Institute of Technology | Mechanical Engineering | NRG |
| Oh, Hae-Soo | M | U of North Carolina-Charlotte | Mathematics | FP |
| Oh, JinHyoung | M | U of Michigan | Aerospace Engineering | NRG |
| Padgett, Clifford | M | North Carolina State U | Materials Science & Engineering | FP |
| Prendergast, Ivy | F | U of Iowa | Applied Mathematics | NRG |
| Reynolds, Daniel | M | Lawrence Livermore National Lab | Center for Applied Scientific Computing | FP |
| Salapaka, Murti | M | Iowa State U | Electrical & Computing Engineering | FP |
| Schall, David | M | North Carolina State U | Materials Science & Engineering | NRG |
| Seelecke, Stefan | M | North Carolina State U | Mechanical & Aerospace Engineering | FP |
| Singler, John | M | ICAM - Virginia Tech | Mathematics | NRG |
| Smith, Ralph | M | North Carolina State U | Mathematics | FP |
| Soberanis, Policarpio | M | U of Iowa | Applied Mathematics | NRG |
| Stanley, Lisa | F | Montana State U | Mathematical Sciences | FP |
| Stockton, John | M | California Institute of Technology | Physics | NRG |
| Taqqu, Murad | M | Boston U and SAMSI | Mathematics and Statistics | FP |
| Tjelmeland, Haakon | M | SAMSI | | FP |
| Toivan, Jari | M | North Carolina State U | CRSC | FP |
| Tran, Hien | M | North Carolina State U | Mathematics | FP |

| | | | | |
|---|---|---|---|---|
| van Handel, Ramon | M | California Institute of Technology | Physics | NRG |
| Vance, Eric | M | Duke U | Statistics | NRG |
| Vannucci, Marina | F | Texas A&M U | Statistics | FP |
| Vidakovic, Brani | M | Georgia Institute of Technology | Industrial & Systems Engineering | FP |
| Vidyashankar, Anand | M | U of Georgia | Statistics | FP |
| Vlahovic, Branislav | M | North Carolina Central U | Physics | FP |
| Vugrin, Eric | M | ICAM - Virginia Tech | Mathematics | NRG |
| Wang, Haonan | M | Colorado State U | Statistics | FP |
| Wang, Kai | M | North Carolina Central U | Physics | FP |
| Wang, Yazhen | M | U of Connecticut | Statistics | FP |
| Wikle, Christopher | M | U of Missouri-Columbia | Statistics | FP |
| Wright, Grady | M | U of Utah | Mathematics | FP |
| Wu, C.F. Jeff | M | Georgia Institute of Technology | Industrial & Systems Engineering | FP |
| Yang, Xingzhou | | North Carolina State U | Mathematics | NRG |
| Zabaras, Nicholas | M | Cornell U | Mechanical & Aerospace Engineering | FP |
| Zheng, Xiaoyu | | U of North Carolina | Mathematics | NRG |
| Zhou, Ruhai | | U of North Carolina | Mathematics | NRG |
| Zietsman, Lizette | F | George Mason U | Mathematical Sciences | FP |

**Education & Outreach Program**
**Two-Day Undergraduate Workshop on Data Mining:**
**Handling the Flood of Data**
NISS-SAMSI Building
*Supported Workshop Participants*
February 13-14, 2004

| Name | Gender | Affiliation | Major/Department | Status |
|---|---|---|---|---|
| Bakewell, Edward | M | U of North Carolina-Wilmington | Mathematics and Statistics | S |
| Barnes, Brad | M | College of Charleston | Computer Science | S |
| Bhat, Anand | M | U of South Florida, Tampa | Mathematics | S |
| Bintu, Lacramioara | F | Brandeis U | Mathematics, Physics and Neuroscience | S |
| Cowie, Nichole | F | George Mason U | Mathematics | S |
| Egoudina, Irina | F | U of Pittsburgh | Statistics | S |
| Fullwood, Michelle | F | Cornell U | Mathematics and Linguistics | S |
| Geraschenko, Anton | M | Brandeis U | Mathematics and Physics | S |
| Gross, Karen | F | U of North Carolina-Wilmington | Mathematics | S |
| Hanson II, Michael | M | U of North Carolina-Wilmington | Mathematics and Statistics | S |
| Nguyen, Belinda | F | U of New Hampshire-Durham | Interdisciplinary Mathematics, Statistics | S |
| Qin, Angie | F | Hunter College | Statistics | S |
| Serang, Gabrielle | F | North Carolina State U | Statistics and Psychology | S |
| Trasti, Jennifer | F | U of North Carolina-Wilmington | Computer Science | S |

**Education & Outreach Program**
**Two-Day Undergraduate Workshop on Data Mining:**
**Handling the Flood of Data**
NISS-SAMSI Building
*Workshop Participants*
February 13-14, 2004

| Name | Gender | Affiliation | Major/Department | Status |
|---|---|---|---|---|
| Bakewell, Edward | M | U of North Carolina-Wilmington | Mathematics and Statistics | S |
| Barnes, Brad | M | College of Charleston | Computer Science | S |
| Bhat, Anand | M | U of South Florida, Tampa | Mathematics | S |
| Bintu, Lacramioara | F | Brandeis U | Mathematics, Physics and Neuroscience | S |
| Budde, Laura Rae | F | North Carolina State U | Statistics | S |
| Cornwell, Travis | M | North Carolina State U | Computer Science and Statistics | S |
| Cowie, Nichole | F | George Mason U | Mathematics | S |
| Egoudina, Irina | F | U of Pittsburgh | Statistics | S |
| Fullwood, Michelle | F | Cornell U | Mathematics and Linguistics | S |
| Geraschenko, Anton | M | Brandeis U | Mathematics and Physics | S |
| Gross, Karen | F | U of North Carolina-Wilmington | Mathematics | S |
| Hanson II, Michael | M | U of North Carolina-Wilmington | Mathematics and Statistics | S |
| Jennings, William | M | U of Rochester | Mathematics | S |
| Kalu, Akuako Patricia | F | Benedict College | Computer Science | S |
| Khan, Nashini | F | Benedict College | Mathematics | S |
| Kipp, Jesse | M | U of Wisconsin-Wilwaukee | Mathematics | S |

| | | | | |
|---|---|---|---|---|
| Nguyen, Belinda | F | U of New Hampshire-Durham | Interdisciplinary Mathematics, Statistics | S |
| Qin, Angie | F | Hunter College | Statistics | S |
| Serang, Gabrielle | F | North Carolina State U | Statistics and Psychology | S |
| Trasti, Jennifer | F | U of North Carolina-Wilmington | Computer Science | S |
| Whitehead, Barron | M | College of Charleston | Pure Mathematics & Computer Science | S |

**Multiscale Model Development and Control Design Program**
**Workshop on Multiscale Challenges in Soft Matter Materials**
Radisson Hotel Research Triangle Park
*Supported Workshop Participants*
February 15-17, 2004

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Bechtel, Stephen | M | Ohio State U | Mechanical Engineering | FP |
| Berlyand, Leonid | M | Penn State U | Mathematics | FP |
| Constantin, Peter | M | The U of Chicago | Mathematics | FP |
| Cook, Pam | F | U of Delaware | Mathematical Sciences | FP |
| Dobrynin, Andrey | M | U of Connecticut | Inst of Materials Sci & Chemical Engineering | FP |
| Fried, Eliot | M | Washington U-St. Louis | Mechanical and Aerospace Engineering | FP |
| Goldsztein, Guillermo | M | Georgia Institute of Technology | Mathematics | FP |
| Goriely, Alain | M | U of Arizona | Mathematics | FP |
| Graham, Michael | M | U of Wisconsin-Madison | Chemical and Biological Engineering | FP |
| Grigoriev, Roman | M | Georgia Institute of Technology | Physics | FP |
| Hosoi, Anette | F | Massachusetts Institute of Technology | Mechanical Engineering | FP |

| | | | | |
|---|---|---|---|---|
| Huang, Zhi-Feng | M | Florida State U | Computational Sci & Information Technology | FP |
| Hyde, E. McKay | M | U of Minnesota | Mathematics | FP |
| Jin, Shi | M | U of Wisconsin-Madison | Mathematics | FP |
| Kilfoil, Maria | F | McGill U | Physics | FP |
| Kumacheva, Eugenia | F | U of Toronto | Chemistry | FP |
| Li, Bo | M | U of Maryland | Mathematics | FP |
| Li, Tiejun | M | Peking U | Mathematical Sciences | FP |
| Mather, Patrick | M | U of Connecticut | Inst of Materials Sci & Chemical Engineering | FP |
| Mucha, Peter | M | Georgia Institute of Technology | Mathematics | FP |
| Nie, Qing | | U of California-Irvine | Mathematics | FP |
| Rey, Alejandro | M | McGill U | Chemical Engineering | FP |
| Rosencrans, Steve | M | Tulane U | Mathematics | FP |
| Shen, Amy | F | Washington U-St. Louis | Mechanical and Aerospace Engineering | FP |
| Tavener, Simon | M | Colorado State U | Mathematics | FP |
| Vukadinovic, Jesenko | M | U of Wisconsin-Madison | Mathematics | FP |
| Wang, Qi | M | Florida State U | Mathematics | FP |
| Wang, Xuefeng | M | Tulane U | Mathematics | FP |
| Yu, Peng | M | Penn State U | Mathematics | FP |
| Zhou, Hong | | U of California-Santa Cruz | Applied Mathematics and Statistics | FP |

**Multiscale Model Development and Control Design Program**
**Workshop on Multiscale Challenges in Soft Matter Materials**
Radisson Hotel Research Triangle Park
*Workshop Participants*
February 15-17, 2004

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Banks, H.T. | M | North Carolina State U | CRSC | FP |
| Bechtel, Stephen | M | Ohio State U | Mechanical Engineering | FP |
| Behringer, Robert | M | Duke U | Physics | FP |
| Berger, Jim | M | SAMSI | | FP |
| Berlyand, Leonid | M | Penn State U | Mathematics | FP |
| Chhajer, Mukesh | M | U of North Carolina | Chemistry | FP |
| Choate, Eric | M | U of North Carolina | Mathematics | NRG |
| Constantin, Peter | M | The U of Chicago | Mathematics | FP |
| Cook, Pam | F | U of Delaware | Mathematical Sciences | FP |
| Cox, Christopher | M | Clemson U | Mathematical Sciences | FP |
| Daniels, Karen | F | Duke U | Physics | FP |
| Dobrynin, Andrey | M | U of Connecticut | Inst of Materials Sci & Chemical Engineering | FP |
| Ermoshkin, Alexander | M | U of North Carolina | Chemistry | FP |
| Forest, Gregory | M | U of North Carolina | Mathematics | NRG |
| Fried, Eliot | M | Washington U-St. Louis | Mechanical and Aerospace Engineering | FP |
| Gartland, Eugene | M | Kent State U | Mathematical Sciences | FP |
| Gleser, Leon | M | U of Pittsburgh | Statistics | FP |

| | | | | |
|---|---|---|---|---|
| Goldsztein, Guillermo | M | Georgia Institute of Technology | Mathematics | FP |
| Goriely, Alain | M | U of Arizona | Mathematics | FP |
| Graham, Michael | M | U of Wisconsin-Madison | Chemical and Biological Engineering | FP |
| Grigoriev, Roman | M | Georgia Institute of Technology | Physics | FP |
| Hosoi, Anette | F | Massachusetts Institute of Technology | Mechanical Engineering | FP |
| Huang, Zhi-Feng | | Florida State U | Computational Sci & Information Technology | FP |
| Hyde, E. McKay | M | U of Minnesota | Mathematics | FP |
| Ji, Chuanshu | M | U of North Carolina | Statistics | FP |
| Jin, Shi | M | U of Wisconsin-Madison | Mathematics | FP |
| Kevrekidis, Yannis | M | Princeton U | Chemical Engineering | FP |
| Kilfoil, Maria | F | McGill U | Physics | FP |
| Krener, Art | M | U of California-Davis | Mathematics | FP |
| Kumacheva, Eugenia | F | U of Toronto | Chemistry | FP |
| Kustanovich, Tamar | F | U of North Carolina | Chemistry | NRG |
| Lada, Emily | F | SAMSI | | FP |
| Lee, Joo Hee | | U of North Carolina | Mathematics | NRG |
| Li, Bo | M | U of Maryland | Mathematics | FP |
| Li, Tiejun | M | Peking U | Mathematical Sciences | FP |
| Liu, Delong | M | NIEHS | | FP |
| Madsen, Louis | M | U of North Carolina | Chemistry and Materials Science | NRG |
| Mancini, Simona Cordier | F | U Paris 6 | Laboratoire J. L. Lions | NRG |

| | | | | |
|---|---|---|---|---|
| Marron, J.S. | M | SAMSI | | FP |
| Mather, Patrick | M | U of Connecticut | Inst of Materials Sci & Chemical Engineering | FP |
| Mattingly, Jonathan | M | Duke U | Mathematics | FP |
| Mucha, Peter | M | Georgia Institute of Technology | Mathematics | FP |
| Mullins, William | M | U.S. Army | Research Office | FP |
| Newell, Andrew | M | North Carolina State U | Mathematics | FP |
| Nie, Qing | | U of California-Irvine | Mathematics | FP |
| Ren, Weiqing | | Princeton U | Mathematics | FP |
| Rey, Alejandro | M | McGill U | Chemical Engineering | FP |
| Rosencrans, Steve | M | Tulane U | Mathematics | FP |
| Rubinstein, Michael | M | U of North Carolina | Chemistry | FP |
| Samulski, E.T. | M | U of North Carolina | Chemistry | FP |
| Sheiko, Sergei | M | U of North Carolina | Chemistry | FP |
| Shen, Amy | F | Washington U-St. Louis | Mechanical and Aerospace Engineering | FP |
| Shusharina, Nadya | F | U of North Carolina | Chemistry | FP |
| Smith, Ralph | M | North Carolina State U | Mathematics | FP |
| Smolka, Linda | F | Duke U | Mathematics | FP |
| Superfine, Richard | M | U of North Carolina | Physics and Astronomy | FP |
| Tavener, Simon | M | Colorado State U | Mathematics | FP |
| Vance, Eric | M | Duke U | Statistics | FP |
| Vukadinovic, Jesenko | M | U of Wisconsin-Madison | Mathematics | FP |

| | | | | |
|---|---|---|---|---|
| Walkington, Noel | M | Carnegie Mellon U | Mathematical Sciences | FP |
| Wang, Qi | M | Florida State U | Mathematics | FP |
| Wang, Xuefeng | M | Tulane U | Mathematics | FP |
| Wang, Zuowei | | U of North Carolina | Chemistry | NRG |
| Waters, Richard | M | U of North Carolina | Mathematics | NRG |
| Yu, Peng | M | Penn State U | Mathematics | FP |
| Zheng, Xiaoyu | | U of North Carolina | Mathematics | NRG |
| Zhou, Hong | F | U of California-Santa Cruz | Applied Mathematics and Statistics | FP |
| Zhou, Ruhai | | U of North Carolina | Mathematics | NRG |

**Planning Workshop for the Random Matrices Program**
American Institute of Mathematics – Palo Alto, CA
*Participants*
February 29-March 1, 2004

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Berger, Jim | M | SAMSI | | FP |
| Bickel, Peter | M | U of California-Berkeley | Statistics | FP |
| Johnstone, Iain | M | Stanford U | Statistics | FP |
| Marron, J.S. | M | SAMSI | | FP |
| McLaughlin, Ken | M | U of North Carolina | Mathematics | FP |
| Nychka, Doug | M | NCAR | | FP |
| Tracy, Craig | M | U of California-Davis | Mathematics | FP |
| Tran, Hien | M | North Carolina State U | Mathematics | FP |

**HOT TOPICS Workshop on Mathematical Sciences Research to Meet National Security Needs**

NISS-SAMSI Building

*Supported Workshop Participants*

April 1-2, 2004

| Name | Gender | Affiliation | Department | Status | Amount Paid |
|---|---|---|---|---|---|
| Aldworth, Jeremy | M | RTI | | FP | fare, room, meals |
| Chang, Harry | M | Army Research Office | | FP | fare, room, meals |
| Cox, Larry | M | National Center for Health Statistics | | FP | fare, room, meals |
| Crowley, Jim | M | SIAM | | FP | fare, room, meals |
| Keller-McNulty, Sallie | F | Los Alamos National Laboratory | | FP | fare, room, meals |
| Kettenring, Jon | M | | | FP | $1,213.41 |
| Launer, Robert | M | Army Research Office | | FP | fare, room, meals |
| Lo, Jim | M | U of Maryland – Baltimore | Mathematics | FP | $461.05 |
| Pollock, Stephen | M | U of Michigan | Industrial Operations | FP | fare, room, meals |
| Robinson, Steve | M | U of Wisconsin | Industrial Engineering | FP | $595.35 |
| Spruill, Nancy | F | Department of Defense | | FP | fare, room, meals |
| Szewcyzk, Bill | M | National Security Agency | | FP | fare, room, meals |
| van den Berg, Eric | M | Telcordia Technologies | | FP | fare, room, meals |
| West, Bruce | M | Army Research Office | | FP | fare, room, meals |

**HOT TOPICS Workshop on Mathematical Sciences Research to Meet National Security Needs**

NISS-SAMSI Building

*Workshop Participants*

April 1-2, 2004

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Aldworth, Jeremy | M | RTI | | FP |
| Banks, David | M | Duke U | Statistics | FP |
| Banks, H.T. | M | NCSU & SAMSI | CRSC | FP |
| Berger, Jim | M | SAMSI | | FP |
| Chang, Harry | M | Army Research Office | | FP |
| Cox, Larry | M | National Center for Health Statistics | | FP |
| Crowley, Jim | M | SIAM | | FP |
| Dinwoodie, Ian | M | Duke U | Statistics | FP |
| Forest, Greg | M | U of North Carolina | Mathematics | FP |
| Goel, Prem | M | Ohio State U | Statistics | FP |
| Jones, Christopher | M | U of North Carolina | Mathematics | FP |
| Karr, Alan | M | NISS | | FP |
| Keller-McNulty, Sallie | F | Los Alamos National Laboratory | | FP |
| Kettenring, Jon | M | | | FP |
| Khattree, Ravi | M | Oakland U | Mathematics & Statistics | FP |
| Launer, Robert | M | Army Research Office | | FP |
| Lin, Xiaodong | M | SAMSI | | NRG |

| | | | | |
|---|---|---|---|---|
| Lo, Jim | M | U of Maryland – Baltimore | Mathematics | FP |
| Marron, J.S. | M | U of North Carolina & SAMSI | Statistics | FP |
| Mattingly, Jonathan | M | Duke U | Mathematics | NRG |
| McLaughlin, Richard | M | U of North Carolina | Mathematics | FP |
| Melton, Tom | M | North Carolina State U | Plant Pathology | FP |
| Pollock, Stephen | M | U of Michigan | Industrial Operations | FP |
| Reiter, Jerome | M | Duke U | Statistics | FP |
| Robinson, Steve | M | U of Wisconsin | Industrial Engineering | FP |
| Shen, Haipeng | M | U of North Carolina | Statistics | NRG |
| Slenning, Barrett | M | North Carolina State U | Farm Animal Health & Resource Management | FP |
| Spruill, Nancy | F | Department of Defense | | FP |
| Szewcyzk, Bill | M | National Security Agency | | FP |
| van den Berg, Eric | M | Telcordia Technologies | | FP |
| West, Bruce | M | Army Research Office | | FP |
| Zhu, Zhengyuan | M | U of North Carolina | Statistics | NRG |

**Multiscale Model Development and Control Design Program**
**Workshop on Fluctuations and Continuum Equations for Granular Flow**
NISS-SAMSI Building
*Supported Workshop Participants*
April 16-17, 2004

| Name | Gender | Affiliation | Department | Status |
|---|---|---|---|---|
| Aronson, Igor | M | Argonne National Laboratory | Materials Science | FP |

| | | | | |
|---|---|---|---|---|
| Berlyand, Leonid | M | Penn State U | Mathematics & Materials Research Inst | FP |
| Clément, Eric | M | U of Paris 6 | | FP |
| Coppersmith, Susan | F | U of Wisconsin | Physics | FP |
| Dahmen, Karin | F | U of Illinois at Urbana-Champaign | Physics | FP |
| Damlamian, Alain | M | U of Paris 6 | Mathematics | FP |
| del Rosario, Ricardo | M | U of the Philippines | Mathematics | NRG |
| Jenkins, James | M | Cornell U | Theoretical and Applied Mechanics | FP |
| Kondic, Lou | M | New Jersey Institute of Technology | Mathematical Sciences | FP |
| Lemaitre, Anael | M | U of California-Santa Barbara | Physics | FP |
| Luding, Stefan | M | Delft U of Technology | DelftChemTech | FP |
| Nielsen, Jorgen | M | Danish Building and Urban Research | | FP |
| Rotter, Michael | M | U of Edinburgh | | FP |
| Shattuck, Mark | M | City College of New York | Physics | FP |
| Sperl, Matthias | M | Technical U of Munich | Physics | FP |

**Multiscale Model Development and Control Design Program**
**Workshop on Fluctuations and Continuum Equations for Granular Flow**
NISS-SAMSI Building
*Workshop Participants*
April 16-17, 2004

| **Name** | **Gender** | **Affiliation** | **Department** | **Status** |
|---|---|---|---|---|
| Aronson, Igor | M | Argonne National Laboratory | Materials Science | FP |
| Banks, H.T. | M | SAMSI & North Carolina State U | Mathematics | FP |

| | | | | |
|---|---|---|---|---|
| Behringer, Robert | M | Duke U | Physics | FP |
| Berlyand, Leonid | M | Penn State U | Mathematics & Materials Research Inst | FP |
| Clément, Eric | M | U of Paris 6 | | FP |
| Coppersmith, Susan | F | U of Wisconsin | Physics | FP |
| Dahmen, Karin | F | U of Illinois at Urbana-Champaign | Physics | FP |
| Damlamian, Alain | M | U of Paris 6 | Mathematics | FP |
| Daniels, Karen | F | Duke U | Physics | FP |
| del Rosario, Ricardo | M | U of the Philippines | Mathematics | NRG |
| Gelfand, Alan | M | Duke U | Statistics | FP |
| Gremaud, Pierre | M | North Carolina State U | Mathematics | FP |
| Jenkins, James | M | Cornell U | Theoretical and Applied Mechanics | FP |
| Ji, Chuanshu | M | U of North Carolina | Statistics | FP |
| Karr, Alan | M | NISS | | FP |
| Kondic, Lou | M | New Jersey Institute of Technology | Mathematical Sciences | FP |
| Lemaitre, Anael | M | U of California-Santa Barbara | Physics | FP |
| Luding, Stefan | M | Delft U of Technology | DelftChemTech | FP |
| Majmudar, Trush | M | Duke U | Physics | NRG |
| Matthews, Matt | M | Duke U | Mathematics | NRG |
| Mattingly, Jonathan | M | Duke U | Mathematics | NRG |
| Nielsen, Jorgen | M | Danish Building and Urban Research | | FP |
| Rotter, Michael | M | U of Edinburgh | | FP |

| | | | | |
|---|---|---|---|---|
| Schaeffer, Dave | M | Duke U | Mathematics | FP |
| Shattuck, Mark | M | City College of New York | Physics | FP |
| Shearer, Michael | M | North Carolina State U | Mathematics | FP |
| Smith, Ralph | M | North Carolina State U | Mathematics | FP |
| Socolar, Josh | M | Duke U | Physics | FP |
| Sperl, Matthias | M | Technical U of Munich | Physics | FP |
| Tighe, Brian | M | Duke U | Physics | NRG |
| Utter, Brian | M | Duke U | Physics | NRG |
| Wambaugh, John | M | Duke U | Physics | NRG |
| Wolpert, Robert | M | Duke U | Statistics | FP |

**APPENDIX E – Workshop Programs and Abstracts**

I. INVERSE PROBLEMS METHODOLOGY IN COMPLEX STOCHASTIC PROBLEMS

*A. Closing Workshop Program*
May 14-15, 2003

**Wednesday – May 14, 2003**
SAMSI-NISS Building

| | |
|---|---|
| **9:15 am** | Opening Remarks |
| **9:30-10:30 am** | Statistics for Science: Perspectives from Pharmacokinetics/Pharmacodynamics<br>**Lewis Sheiner**, University of California, San Francisco |
| **10:30-11:30 am** | A Simulation Based Comparison Between Parametric and Semiparametric Methods in a PBPK Model<br>**Yanyuan Ma**, SAMSI |
| **11:30-11:45 am** | Discussions |
| **11:45-1:00 pm** | Lunch |
| **1:00-2:00 pm** | Non-Stationary Inverse Problems and Dynamical Prior Models<br>**Erkki Somersalo**, Helsinki University of Technology, Finland |
| **2:15-3:00 pm** | Inverse Problems in Complex Model Validation<br>**Danny Walsh**, SAMSI |
| **3:00-3:30 pm** | Discussions |

**Thursday – May 15, 2003**
SAMSI-NISS Building

| | |
|---|---|
| **9:30-10:30 am** | A Hierarchical Bayesian Spatio-Temporal Model for Predicting the Spread of Invasive Species Given Uncertain Observations<br>**Christopher K. Wikle**, University of Missouri-Columbia |
| **10:45-11:30 am** | 2D Electromagnetic Parameter Identification for a Debye Polarization Model<br>**Johnathan Bardsley**, SAMSI |
| **11:30-11:45 am** | Discussions |
| **11:45-1:00 pm** | Lunch |
| **1:00-2:00 pm** | Iterative Regularization of Nonlinear Inverse Problems: Deterministic Convergence Theory, Ideas on Incorporating Uncertainty<br>**Heinz Engl**, Johannes Kepler Universitat, Linz |

**2:00-3:30 pm** Closing Discussions and Remarks

II. LARGE SCALE COMPUTER MODELS FOR ENVIRONMENTAL SYSTEMS

*A. One-day Workshop In Porous Media Program*

**Friday – May 16, 2003**
SAMSI-NISS Building

| | |
|---|---|
| **9:00-9:45 am** | Closure Issues<br>**Greg Forest and Roberto Camassa** |
| **9:45-10:30 am** | Theory and APS experiments<br>**Katherine Culligan** |
| **10:30-11:00 am** | Break |
| **11:00-11:45 am** | Image Processing<br>**Dorthe Wildenschild** |
| **11:45-12:30 pm** | Lattice Boltzmann Simulation<br>**Doris Pan** |
| **12:30-1:30 pm** | Lunch |
| **1:30-2:15 pm** | 2D Micromodels<br>**Laura Pyrak-Nolte** |
| **2:15-3:00 pm** | Pore Drainage<br>**Markus Hilpert** |
| **3:00-3:30 pm** | Break |
| **3:30-4:15 pm** | Level Set Methods<br>**David Adalsteinsson** |
| **4:15-5:00 pm** | Group - Plans for Future |

*B. Program for Workshop on Spatio-Temporal Modeling* (in association with the National Center for Atmospheric Research's Geophysical Statistics Project)
June 1-6, 2003 Boulder, CO

**Sunday – June 1, 2003**
Foothills Lab 2, Room 1022

**7:00-9:00 pm** Mixer at the Millennium Hotel (Garden area)

**Monday – June 2, 2003**
Foothills Lab 2, Room 1022

| | |
|---|---|
| **8:45-10:00 am** | Numerical Models of Regional Air Quality: (Lining Up Model and Measurements to Probe and Address Uncertainty)<br>**Robin Dennis**, NOAA/Environmental Protection Agency |
| **10:00-10:15 am** | Break |
| **10:15-11:30 am** | Variance-Covariance Modeling and Estimation for Multi-Resolution Spatial Models<br>**Noel Cressie**, The Ohio State University |
| **11:30-1:00 pm** | Lunch & Poster Session |
| **1:00-2:15 pm** | Introduction to Methods for Areal (Lattice) Data<br>**Brad Carlin**, University of Minnesota |
| **2:15-3:15 pm** | Controlling for confounding in time series studies of air pollution and mortality: How smooth (or rough) should we be?<br>**Francesca Dominici**, John Hopkins University |
| **3:15-3:30 pm** | Break |
| **3:30-4:30 pm** | Disease Mapping with Disparate Spatial Data<br>**Robert Wolpert**, Duke University |
| **4:30-5:30 pm** | Geostatistical Modeling<br>**Richard Smith**, University of North Carolina |

**Tuesday – June 3, 2003**
Foothills Lab 2, Room 1022

| | |
|---|---|
| **8:45-10:00 am** | Bayesian Perspective on Inference for Space-Time Processes<br>**Mark Berliner**, The Ohio State University |
| **10:00-10:15 am** | Break |
| **10:15-11:30 am** | Incorporating Scientific Priors in Hierarchical Spatio-Temporal Models: An Invasive Species Case Study<br>**Chris Wikle**, University of Missouri |
| **11:30-1:00 pm** | Lunch & Poster Session |

**1:00-2:15 pm** □Ensemble Filters for Atmosphere and Ocean Data Assimilation□
**Jeff Anderson**, Geophysical Fluids Dynamics Laboratory

**2:15-3:15 pm** □Spatial-Temporal Aspects of Water Quality□and □Spatial Ensemble Estimates of Temporal Trends in Acid Neutralizing Capacity□
**N. Scott Urquhart** & **F. Jay Breidt**, Colorado State University

**3:15-3:30 pm** Break

**3:30-4:30 pm** □Measuring the information content of ensemble predictions in dynamical systems relevant to atmosphere and ocean□
**Richard Kleeman**, Courant Institute of Mathematical Sciences, New York University

**4:30-5:30 pm** □Review of Spectral Methods for Spatial Processes□
**Montserrat Fuentes**, North Carolina State University

**Wednesday – June 4, 2003**
Foothills Lab 2, Room 1022

**8:45-11:30 am** □Detection of Anthropogenic Climate Change□
**Gabriele Hegerl**, Duke University

**Ben Santer**, Lawrence Livermore National Laboratory

□Deconstructing Recent Global-Mean Temperature Changes□
**Tom M. L. Wigley**, National Center for Atmospheric Research

(15 minute break at 10:15)

**11:30-1:00 pm** Lunch & Poster Session

**1:00-2:15 pm** □Quantifying Uncertainty for Non-Gaussian Ensembles in Climate Prediction□
**Andrew J. Majda**, Courant Institute of Mathematical Sciences, New York University

**2:15-3:15 pm** □Homogenization of Gravity Currents in Heterogeneous Porous Media□

**Richard McLaughlin**, University of North Carolina

**3:15 pm** Free Time

**Thursday – June 5, 2003**
Foothills Lab 2, Room 1022

**8:45-10:00 am** "Environmental Problems, Spatial Modeling, and Bayesian Inference"
**Alan Gelfand**, Duke University

**10:00-10:15 am** Break

**10:15-11:30 am** "The Matrix Revisited: Spatial Analysis of Large Data Sets"
**Doug Nychka**, National Center for Atmospheric Research

**11:30-1:00 pm** Lunch

**1:00-2:15 pm** "Scientific air pollution mapping across space and time: Dealing with data uncertainties and the integration of physical laws."
**Marc Serre** & **George Christakos**, University of North Carolina

**2:15-3:15 pm** "Models for Spatial-Temporal Covariances"
**Michael Stein**, University of Chicago

**3:15-3:30 pm** Break

**3:30-4:30 pm** "Predictive spatio-temporal models for spatially sparse environmental data" & "Skew-elliptical distributions for environmental data"
**Marc Genton**, North Carolina State University

**Friday – June 6, 2003**
Foothills Lab 2, Room 1022

**8:30-12:00 pm** Working groups

*C. Workshop on Spatio-Temporal Modeling Abstracts*

**Jeffrey Anderson**
Geophysical Fluid Dynamics Laboratory

"Ensemble Filters for Atmosphere and Ocean Data Assimilation"

Modern data assimilation for the atmosphere and ocean combines information from observations and a prediction model to produce an estimate of the state of the physical system. Data assimilation is used not only to generate 'analyses' of the system state, but also to produce initial conditions for prediction models. It can also be used to improve both the observing system and the prediction models.

Ensemble filtering algorithms for data assimilation have recently begun to mount a challenge to variational methods as candidates for next generation prediction systems. A simple derivation of several varieties of ensemble filters as Monte Carlo approximations to the solutions of a non-linear filtering problem is presented. Most of the ensemble methods in use can be derived as a one-dimensional filter followed by a sequence of linear regressions. Selected results from applications in both low-order models and large operational prediction models demonstrate the power of these methods.

**L. Mark Berliner**
Ohio State University
Department of Statistics

Bayesian perspective on inference for space-time processes

I begin with a very brief review of selected issues and approaches to space-time statistical analysis. I then transition through three basic motivations of the Bayesian perspective. The first of these involves clarification of the Bayesian interpretation of some common procedures (e.g., the Kalman filter). The second is the traditional Bayesian method of endowing model parameters with prior distributions and applying Bayes' Theorem. The third component of the discussion is the development of Bayesian models which actively rely on physical reasoning. To motivate these physical-statistical approaches, a brief review of stochastic modeling of climate and weather processes in presented. Selected examples are presented.

**Brad Carlin**
University of Minnesota

Introduction to Methods for Areal (Lattice) Data

We present a review of both exploratory tools and modeling approaches which are customarily applied to data collected for areal units. We have in mind general, possibly irregular geographic units (as for example are common in spatial disease mapping) but of course include the special case of regular grids of cells (pixels). After a brief summary of common exploratory tools (such as Moran's I and Geary's C), we proceed on to a development of various results in Markov random field theory that underlie spatial lattice modeling, such as Brook's Lemma. Conditionally autoregressive (CAR), intrinsically autoregressive (IAR), and simultaneously autoregressive (SAR) models are also described and compared. We then outline use of these models in spatial disease mapping, highlighting computational issues, especially those related to model identifiability and use of the WinBUGS software. If time permits, we will also highlight multivariate generalizations of these models (e.g., the so-called MCAR model) and further related application areas (e.g., the use of CAR and and MCAR models in spatial and spatio-temporal frailty modeling).

**Noel Cressie**
Ohio State University
Department of Statistics

"Dynamic Multi-Resolution Spatial Models"

The material presented in this talk is the result of joint research by Gardar Johannesson (Lawrence Livermore), Noel Cressie (Ohio State) and Hsin-Cheng Huang (Academia Sinica). We consider the problem of spatial-temporal prediction of global processes using a model that recognizes multiple resolutions in the spatial domain. Here, optimal spatial-prediction procedures can be shown to be extremely fast. Similar ideas can be used in the spatial-temporal domain; a vector autoregressive model is assumed at the coarsest resolution and, at each time-point, a multi-resolution spatial strucure is modeled. Then the idea is to use Bayesian updating to make the prior distribution of the coarse-resolution process more informative as time proceeds. Our spatial-temporal methodology will be compared to the spatial-only methodology on data from the Total Ozone Mapping Spectrometer (TOMS) instrument, on the Nimbus-7 satellite.

**Robin Dennis**
Environmental Protection Agency
NOAA

"Numerical Models of Regional Air Quality: (Lining Up Model and Measurements to Probe and Address Uncertainty)"

The basic paradigm of a numerical model for regional air quality will be briefly described and the main objectives of air quality models for EPA will be noted. Two key conceptual models of scientific processes being simulated (ozone and inorganic fine particles) underlying many of the predictions of interest will be briefly described. A general perspective on uncertainties, from an air quality modeling viewpoint, separated into transient (random) and persistent (systematic) uncertainties will be illustrated with examples. Then four areas for potential collaboration will be discussed: Persistent errors and model evaluation, reducing (hopefully) uncertainty in inputs through inverse modeling, combining models and measurements through spatial-temporal modeling, and probing process relationships through multivariate analysis. Issues of model performance uncovered in highly resolved data turn into persistent errors in more typical spatial data sets that have longer integration times. Some of the sources of persistent error will be noted and the fact they can change with season will be illustrated. The importance of input errors will be illustrated with the example of ammonia emissions and the inorganic fine particle system. The use of inverse modeling to try to address input errors is presented as a valuable tool to be further developed. The spatial exploration of input errors is an important need, particularly for the health community. This user community desire for space-time interpolation that combines model and measurements will be discussed.

**Francesca Dominici**
John Hopkins University

"Controlling for confounding in time series studies of air pollution and mortality: How smooth (or rough) should we be?"

Time series studies of air pollution and health are aimed at estimating associations between day-to-day variations in air pollution and day-to-day variations in daily mortality counts in the presence of: 1) observed time-varying confounders (e.g., weather); and 2) time-varying unobserved confounders such as respiratory influenza and trends in survival. To eliminate long-term trends and seasonal variations in the mortality time series, a smooth function of calendar time, f(t), is included in the regression formulation. The statistical\epidemiological target is to determine the degree of smoothness of f(t) that sufficiently reduces confounding bias when estimating the pollution coefficient.

The choice of the number of degrees of freedom (df) used to represent f(t) is one of the most discussed statistical issues in time series analyses of air pollution and health. This choice is critical because it determines the time scales of variations in the health outcome and exposure used for the estimation of the air pollution coefficient. Choosing too small a df, that is over-smoothing f(t), might result in confounding bias. At the other extreme, choosing too large a df, that is under-smoothing f(t), might wash out the pollution effect by over adjusting or inflating the statistical variance of the pollution coefficient estimate.
Current approaches for df-selection in environmental epidemiology are: data-driven methods, that is the number of degrees of freedom in the smooth function of time is estimated based upon optimality criteria such as the Akaike Information Criteria, or based upon prior choices on df supported by sensitivity analyses.

In this talk we show that data-driven methods for df-selection are generally not suitable strategies for removing confounding bias. We then introduce a Bayesian Model Averaging approach for estimating the pollution coefficient which takes into account prior information and uncertainty about df, that is about the time scales of variations in the time series where confounding might occur. Methods are applied to time series data from the NMMAPS study.

This is joint work with Thomas A. Louis, Aidan McDermott, and Jonathan M. Samet.

**Montserrat Fuentes**
North Carolina State University
Department of Statistics

Review of spectral methods for spatial processes

Spectral methods are a powerful tool for studying the spatial structure of random fields and generally offer significant computational benefits. The objective of this talk is to introduce the audience to Fourier analysis and spectral methods for spatial temporal processes. I will review discrete and continuous Fourier analysis, Fourier representation of non-periodic functions, the aliasing phenomenon in the spectral domain, the Fast Fourier Transform (FFT) to efficiently approximate Fourier integrals, and some of the spectral methods to approximate a Gaussian likelihood.

A stationary process has a spectral representation in terms of sine and cosine waves, we will generalize this result to nonstationary and nonseparable spatial temporal processes.

I will present some of the commonly used classes of spectral densities for stationary processes, and some new one models for nonstationary and nonseparable space-time processes.

All these methods will be applied to air pollution data provided by the US EPA.

**Alan E. Gelfand**
Duke University
Institute of Statistics and Decision Sciences

Environmental Problems, Spatial Modeling and Bayesian Inference

The goal of this presentation is to consider several environmental problems which can be usefully addressed using spatial process models and to detail the implementation of and benefits of fitting such models within a Bayesian framework. First, we review the formulation of general classes of hierarchical spatial models introducing spatial random effects modeled through spatial processes. We then briefly discuss simulation based strategies for the fitting of such models.

We then consider four illustrative problems as follows:

(1) multivariate modeling of pollution surfaces using coregionalization
(2) spatio-temporal analysis of pollution surfaces using dynamic models with spatially varying coefficients
(3) detecting gradients in pollution surfaces using directional derivative processes
(4) population-adjusted toxin exposure analysis using covariate-weighted spatial CDF's.

**Marc G. Genton**
North Carolina State University
Department of Statistics

Predictive spatio-temporal models for spatially sparse environmental data

We present a family of spatio-temporal models which are geared to provide time-forward predictions in environmental applications where data is spatially sparse but temporally dense. That is measurements are made at few spatial locations (stations), but at many regular time intervals. When predictions in the time direction is the purpose of the analysis, then spatial-stationarity assumptions which are commonly used in spatial modeling, are not necessary. The family of models proposed does not make such assumptions and consists in a vector autoregressive (VAR) specification, where there are as many time series as stations. However, by taking into account the spatial dependence structure, a model building strategy is introduced, which borrows its simplicity from the Box-Jenkins strategy for univariate autoregressive (AR) models for time series. As for AR models, model building may be performed either by displaying sample partial correlation functions, or by minimizing an information criterion. Two environmental data sets are studied. In particular, we find evidence that a parametric modeling of the spatio-temporal correlation function is not appropriate because it rests on too strong assumptions. Moreover, we propose to compare model selection strategies with an out-of-sample validation method based on recursive prediction errors.

This is joint work with Xavier de Luna, Umea University.

**Marc G. Genton**
North Carolina State University
Department of Statistics

Skew-elliptical distributions for environmental data

We describe a flexible family of multivariate skew-elliptical distributions that can account for skewness, heavy tails, and multimodality in the data. It is straightforward to simulate from this family which possesses several invariance properties. In particular, we show that the distribution of the sample autocovariance function in time series and of the sample variogram in spatial statistics do not depend on the skewness of these distributions. We then discuss the use of a certain type of skew-normal distributions in the Kalman filter. This later topic is ongoing work with Philippe Naveau.

**Gabriele Hegerl**
Duke University
Nicholas School of the Environment

Detection of Anthropogenic Climate Change

The Intergovernmental Panel on Climate Change concluded in 2001 report that "most of the warming observed over the last 50 years is likely due to the increase in greenhouse gases". The evidence leading to this conclusion will be briefly reviewed, including recent results further addressing uncertainties. For example, reconstructions of hemispheric scale surface temperature over the last millennium can help to address uncertainties in estimates of natural (internal and externally influenced) climate variability. The comparison between simulations with a simple climate model and these data suggest that the probability density function for the expected temperature increase due to CO2 doubling can be estimated based on comparisons between simulations with a simple climate model and this long temperature record. While detection of climate change in large scale temperature indicators is an important exercise in model validation, society is more concerned about changes in variables that affect it more directly, such as changes in rainfall and climate extremes. The study of changes in extreme events raises interesting statistical questions, such as the relationship between point observations and area averaged rainfall data. The outlook for detection of changes in temperature and precipitation extremes is discussed by studying changes from simulations with two different atmosphere-ocean general circulation models.

**Richard Kleeman**
New York University
Courant Institute of Mathematical Sciences

Measuring the information content of ensemble predictions in dynamical systems relevant to atmosphere and ocean

In the past two years a new theoretical framework for analyzing predictive information has been developed by the author for geophysical applications. This has been motivated by the very practical issue of determining how predictive utility varies from one prediction to another. The framework relies on entropic information theoretic functionals on probability distribution functions (pdfs). The latter describe the random variables of interest. In practical contexts only an estimate of these functions is possible usually by means of a Monte-Carlo ensemble or sample of predictions which are distributed according to the (unknown) pdf. Obviously this imperfect knowledge of how the random variables are distributed implies a further reduction in the

information content of the prediction made. We use Bayesian analysis tools to quantify this loss and discuss the relation to maximum entropy methodologies from mathematical statistics.

**Andrew J. Majda**
New York University
Courant Institute of Mathematical Sciences

"Quantifying Uncertainty for Non-Gaussian Ensembles in Climate Prediction"

Many situations in complex systems require quantitative estimates of the lack of information in one probability distribution relative to another. In short term climate and weather prediction, examples of these issues might involve the lack of information in the historical climate record compared with an ensemble prediction, or the lack of information in a particular Gaussian ensemble prediction strategy involving the first and second moments compared with the non-Gaussian ensemble itself. The relative entropy is a natural way to quantify this information. Here a recently developed mathematical theory for quantifying this lack of information is converted into a practical algorithmic tool. The theory involves explicit estimators obtained through convex optimization, principal predictability components, a signal/dispersion decomposition, etc. An explicit computationally feasible family of estimators is developed here for estimating the relative entropy over a large dimensional family of variable through a simple hierarchical strategy. Many facets of this computational strategy for estimating uncertainty are applied here for ensemble predictions for two "toy" climate models developed recently: the Galerkin truncation of the Burgers-Hopf equation and the Lorenz '96 model.

This is a joint work with Rafail Abramov of the Courant Institute of Mathematical Sciences at New York University.

**Richard M. McLaughlin**
University of North Carolina
Department of Mathematics

"Homogenization of Gravity Currents in Heterogeneous Porous Media"

We will overview homogenization methods context of computing effective mixing coefficients in several contexts. First, we examine the homogenized averaging of a passive scalar equation diffusing in the presence of a prescribed, variable coefficient fluid flow. Then we turn to the problem of assessing the effect of heterogeneity upon the motion of a slumping gravity current in porous media. The lecture is planned to motivate the need for good closure schemes, while offering an introduction to homogenization methods.

**Doug Nychka**
National Center for Atmospheric Research
Geophysical Statistics Project

"The Matrix Revisited: Spatial Analysis of Large Data Sets"

A spatial analysis often involves manipulating covariance matrices and solving linear systems. This talk present several computational and modeling strategies for dealing with the matrix

calculations when spatial data sets are large. Some of these approaches include iterative methods for solving large linear systems and inducing sparsity in the covariance matrix.

**Marc L. Serre and George Christakos**
University of North Carolina
Department of Environmental Sciences and Engineering

Scientific air pollution mapping across space and time: Dealing with data uncertainties and the integration of physical laws.

The study of the spatiotemporal distribution of air pollutants is an important issue due to the health risks associated with these pollutants. In the U.S., criteria air pollutants are measured throughout the country by means of an extensive network of monitoring stations. The high variability of air pollutants across space and time and the varying levels of data accuracy introduce major sources of uncertainty in the study of their spatiotemporal distribution. The last two decades the BME modelling has offered a powerful epistemic framework for integrating various knowledge bases (physical laws, scientific theories, primitive equations, uncertain data sources, secondary information etc.) and producing realistic pictures of air pollutant distribution in a composite space-time domain. In this work we use BME to integrate physical knowledge bases about air pollution variability and to map efficiently the annual arithmetic averages of particulate matter (PM) across the U.S. during the 1984-2000 time period. BME rigorously processes probabilistic (soft) data describing different accuracy levels in the PM measurements and produces informative representations of the spatiotemporal air pollutant distribution and its associated mapping uncertainty. Several applications can be found in the literature in which BME has provided a mathematically rigorously and physically meaningful framework for integrating several kinds of physical laws (in the form of partial differential equations, algebraic equations etc.). This work presents recent developments involving an advection-dominated air pollutant transport equation in space-time. It is shown that the integration of the specified physical law incorporates valuable knowledge about wind and pollution sources in the mapping process and leads to improved predictions of air pollution concentrations.
Key words: Spatiotemporal, BME, random fields, air pollution, particulate matter

**Richard L. Smith**
University of North Carolina
Department of Statistics

Geostatistical Modeling

Geostatistical modeling refers to methods of spatial and spatial-temporal statistics based on fitting a suitable covariance function (or variogram) to the data. They are usually the preferred method of analysis for data collected at point locations (as opposed to areal averages) and when these locations are not arranged on a regular lattice, as is typical of most environmental monitoring data. In this talk I will review the main steps of the methodology, in particular (a) defining models for spatial covariances and variograms, (b) model identification and parameter estimation, (c) prediction and interpolation ("kriging" and its extensions), (d) some of the simpler extensions to spatial-temporal models (the separable and "repeated measures" model). As an application, I shall consider the problem of estimating mean levels of fine particulate matter across a three-state region, to assess compliance with EPA air pollution standards.

**Michael Stein**
University of Chicago
Department of Statistics

"Models for Spatial-Temporal Covariances"

A good model for the covariance function of a stationary process in space and time should accurately describe the variances and correlations of all linear combinations of the process. In particular, it does not suffice to find a model that describes the purely temporal covariances and the purely spatial covariances accurately. Rather, it is critical to capture the spatial-temporal interactions as well. We consider a number of properties of spatial-temporal covariance functions and how these relate to the spatial-temporal interactions of the process. First, we examine how the smoothness away from the origin of a spatial-temporal covariance function affects, for example, temporal correlations of spatial differences. Second, we examine the implications of a Markov assumption in time on spatial-temporal covariance functions. Third, we consider models that are asymmetric in space-time: the correlation between site A at time t and site B at time s is different than the correlation between site A at time s and site B at time t. We examine some of these issues for sulfate concentrations as measured by monitors, as produced by an air quality model (CMAQ) and for the difference between observed and modeled concentrations.

This is a joint work with Mikyoung Jun.

**N. Scott Urquhart** and **F. Jay Breidt**
Colorado State University
Department of Statistics

"Spatial-Temporal Aspects of Water Quality"

Environmental agencies, federal, state and tribal, must evaluate water quality using chemical, biological, thermal and usage criteria over vast spatial and long-term temporal domains. The Environmental Protection Agency (EPA) is publishing a report on the possible effects of the Clean Air Act Amendments of 1990 on the acid status of surface waters in several regions of the US. The first part of this talk will identify statistically important features of the data underlying this report, and limitations of statistical techniques available to design such studies and analyze the resulting data.

The second part of the talk will describe a specific estimation problem for acid status of surface waters in the Northeastern United States. One of the tools used to evaluate characteristics of acidity is the cumulative distribution function of slope trends (acid concentration/year). An understanding of the distribution of these slopes helps evaluate the impact of the Clean Air Act Amendments of 1990. For example, the proportion of lakes whose acidic concentration has been decreasing can be estimated. A hierarchical model is constructed to describe these slopes as functions of available auxiliary information, and constrained Bayes techniques are used to estimate the ensemble of slope values.

**Chris Wikle**
University of Missouri
Department of Statistics

Incorporating Scientific Priors in Hierarchical Spatio-Temporal Models: An Invasive Species Case Study

There is increasing interest in predicting ecological processes. Methods to accomplish such predictions must account for uncertainties in observation, sampling, models, and parameters. Statistical methods for spatio-temporal processes are powerful, yet difficult to implement in complicated, high-dimensional settings. However, recent advances in hierarchical Bayesian formulations for such processes can be utilized for ecological prediction. These formulations are able to account for the various sources of uncertainty, and can incorporate scientific judgment in a probabilistically consistent manner. For example, analytical diffusion models can serve as motivation for the hierarchical model for invasive species. We demonstrate by example that such a framework can be utilized to predict spatially and temporally, population characteristics of birds from the Breeding Bird Survey. Time permitting, alternative specifications of underlying process will be considered.

**Robert L. Wolpert**
Duke University
Institute of Statistics and Decision Sciences

Disease Mapping with Disparate Spatial Data

Ecological regression studies are widely used to explore relationships between disease rates and levels of exposure to environmental risk factors. The raw data for such studies, such as disease case counts, environmental pollution concentration measurements and the reference population at risk, are measured at disparate levels of spatial aggregation but are commonly accumulated to a single common geographical scale to facilitate statistical analysis. In this traditional approach heterogeneous exposure distributions within the aggregate areas may lead to biased inference, while individual attributes such as age, gender and smoking habits must either be summarized to provide area level covariate values or used to stratify the analysis.

This presentation offers a spatial regression analysis of the effect of traffic pollution on respiratory disorders in children. The analysis features data measured at disparate, non-nested scales, including spatially varying covariates, latent spatially varying risk factors, and case-specific individual attributes.

The problem of disparate discretizations is overcome by relating all spatially-varying quantities to a continuous latent underlying random field. Case-specific individual attributes are accommodated by treating cases as a marked point process. Inference in these hierarchical Poisson/L'evy models is based on simulated samples drawn from Bayesian posterior distributions, using Markov chain Monte Carlo methods.

## *D. Workshop on Spatio-Temporal Modeling Poster Session List*

**Monday – June 2, 2003**

- **Catherine Calder**, Duke University
  Exploring Latent Structure in Multivariate Spatial Temporal Process using Process Convolutions

- **Marco Ferreira**, Universidade Federal de Rio de Janeiro
  Bayesian Inference for Proper Gaussian Markov Random Fields

- **John Holt**, University of Guelph
  Multiple testing in environmental epidemiology A case study and Spatial and temporal modeling of sleeping sickness in South-East Uganda

- **Isin Ozaksoy**, North Carolina State University
  Evaluation of the seismic activities of Turkey

- **Chris Paciorek**, Carnegie Mellon University
  A class of convolution-based nonstationary covariance functions

- **Jun Zhu**, University of Wisconsin
  A Multiresolution Tree-Structure Spatial Linear Model

- **Zhengyuan Zhu**, University of North Carolina
  Optimal Sampling Design for Gaussian Random Fields

**Tuesday – June 3, 2003**

- **Li Chen**, North Carolina State University
  Spatial-temporal modeling for wind data
  (joint with M. Fuentes and J. M. Davis, North Carolina State University)

- **Dana Draghicescu**, University of Chicago
  A model for Spatio-temporal prediction of ground-level ozone mixing ratios (joint work with V. Dukic, G. Eshel, J. Frederick, E. Naureckas, P. Rathouz, M. Stein, and A. Zubrow, University of Chicago)

- **Yulia Gel**, University of Washington
  Accessing uncertainty in numerical mesoscale weather prediction via ensembles of forecasts (joint with A. Raferty and T. Gneiting, University of Washington)

- **Serge Guillas**, University of Chicago
  Space-time modeling of stratospheric ozone

- **Jaechoul Lee**, University of Georgia
  Trends in United States Temperature Extremes

- **Anders Malmberg**, University of Lund
  A real-time assimilation model for near-surface ocean winds

- **Jonathan Stroud**, University of Pennsylvania
  Space-time Modeling of Mexico City Ozone Levels
  (joint with G. Huerta, University of New Mexico and B. Sanso, University of California at Santa Cruz)

**Wednesday – June 4, 2003**

- **Daniel Fink** and **Marty Wells**, Cornell University
  "Adaptive Multi-Order Penalized Splines"

- **Sandra McBride**, Duke University
  "Hierarchical Bayesian Calibration: An Application to Ambient Air Pollution Measurement Data"

- **Bruno Sanso**, University of California-Santa Cruz
  "Spatio-temporal models based on discrete convolutions" (joint with A. Schmidt)

## III. CHALLENGES IN STOCHASTIC COMPUTATION

### *A. Closing Workshop Program*
June 26-28, 2003

**Thursday – June 26, 2003**
Radisson Governors Inn

**8:00 am** Continental Breakfast

**8:15-8:45 am** Registration Check-In

**8:45 am** Welcome and Introduction

**9:00-10:30 am** **Session 1** - Graphical Models A
*Chair:* Chris Hans, Duke University

"MCMC Methods for Bayesian Data Mining"
**Paulo Giudici**, University of Pavia

"Graphical Gaussian Model Selection in a Bayesian Framework"
**Helene Massam**, York University

**10:30-11:00 am** Coffee break

**11:00-12:30 pm** **Session 2** - Graphical Models B
*Chair:* Carlos Carvalho, Duke University

"Adventures in Graph Space: Practical Issues in Exploring Graphical Structures for Moderate to Moderately Large Numbers of Variables"
**Beatrix Jones**, SAMSI and Duke University

"Compositional Regressions, DAGs and Fitting High-Dimensional Gaussian Graphical Models"
**Adrian Dobra**, SAMSI and Duke University

**12:30-1:30 pm** Lunch

**1:30-3:00 pm** **Session 3**
*Chair:* Chris Carter, Duke University & SAMSI

Exact MCMC Goodness of Fit Tests
**Julian Besag**, University of Washington

Stochastic Simulation Methods for the Number of Components in a Mixture
**Sujit Sahu**, University of Southampton

**3:00-3:30 pm** Coffee Break

**3:30-5:00 pm** **Session 4** - Model Selection A
*Chair:* Joe Ibrahim, University of North Carolina

Mixtures of G-Priors and Variable Selection
**Feng Liang**, Duke University

Marginal Likelihoods and Bayes Factors
**Rui Paulo**, SAMSI & NISS

**Friday – June 27, 2003**
Radisson Governors Inn

**8:00 am** Continental Breakfast

**8:30-9:00 am** Registration Check-In

**9:00-10:30 am** **Session 5** - Model Selection B
*Chair:* M.J. Bayarri, University of Valencia

Nonparametric Regression Variable Selection COSSO
**Helen Zhang**, North Carolina State University

Stochastic Search via Sampling Adaptively without Replacement
**Merlise Clyde**, Duke University

**10:30-11:00 am** Coffee break

**11:00-12:30 pm** **Session 6** - Contingency Tables & Related Topics
*Chair:* Ian Dinwoodie, Tulane University & Duke SAMSI Fellow

Markov Chain Moves for Generating Contingency Tables with Fixed Weighted Row Sums
**Mark Huber**, Duke University

Sequential Importance Sampling for Generating Contingency Tables with Fixed Weighted Row Sums
**Yuguo Chen**, Duke University

**12:30-1:30 pm** Lunch

**1:30-3:00 pm** **Session 7** - Contingency Tables & Related Topics A
*Chair:* Mark Huber, Duke University

"Making Inferences from Arbitrary Sets of Conditionals and Marginals for Contingency Tables"
**Aleksandra Slavkovic**, Carnegie Mellon University

"Algebraic Geometry of Bayesian Networks with Hidden Variables"
**Luis David Garcia**, Virginia Tech

**3:00-3:30 pm** Coffee Break

**3:30-5:00 pm** **Session 8** - Contingency Tables & Related Topics B
*Chair:* Michael Nicholas, Duke University

"Variations on Barvinok's Counting Algorithm and Applications to Multiway Contingency Tables"
**Ruriko Yoshida**, University of California, Davis

"Markov Bases of Binary Graph Models"
**Seth Sullivant**, University of California, Berkeley

**Saturday – June 28, 2003**
SAMSI-NISS Building

**8:00 am** Continental Breakfast

**9:00-10:30 am** **Session 9** - Financial Models A
*Chair:* Yuguo Chen, Duke University

"Gaussian Approximations in MCMC Computation of Option Prices with Stochastic Volatility Models"
**Chuanshu Ji**, University of North Carolina

"Multi-scale Stochastic Volatility"
**Jean-Pierre Fouque**, North Carolina State University

**10:30-11:00 am** Coffee break

**11:00-12:30 pm** **Session 10** - Financial Models B
*Chair:* Chuanshu Ji, University of North Carolina

"MCMC for Estimation of Multiscale Stochastic Volatility Models"
**German Molina**, Duke University

"Pricing Asian Options and Variance Swaps with Volatility Scales"
**Sean Han**, North Carolina State University

**12:30-1:30 pm** Lunch

**1:30-3:00 pm** **Session 11 -** Financial Models and Time Series
*Chair:* Jean-Pierre Fouque, North Carolina State University

"A Stochastic Computational Method for Portfolio Optimization Problems"
**Tao Pang**, North Carolina State University

"Bayesian Analysis of Random Coefficient Autoregressive Models"
**Sujit Ghosh**, North Carolina State University

**3:00-3:30 pm** Coffee Break & Close

**6:00 pm** STOCOM & DUKE-SAMSI SOCIAL EVENT
Transportation between SAMSI and the Radisson Governor's Inn provided by Carolina Livery until 10:00pm.

## *B. Closing Workshop Abstracts*

**Julian Besag**
University of Washington
Department of Statistics

"Exact MCMC Goodness of Fit Tests"

Goodness-of-fit tests can be useful as an exploratory tool in identifying parsimonious statistical descriptions of datasets. In ordinary Monte Carlo tests, the data are compared with random samples generated from the proposed formulation, possibly after appropriate conditioning to eliminate parameters. When direct simulation is impracticable, it can be replaced by Markov chain Monte Carlo, whilst preserving the exactness of the p-value. The talk will describe and illustrate the procedures and mention some open problems.

**Yuguo Chen**
Duke University
Institute of Statistics & Decision Sciences

"Sequential Importance Sampling for Generating Contingency Tables with Fixed Weighted Row Sums"

Motivated by the exact conditional inference problems for contingency tables with ordered categories, we describe a sequential importance sampling strategy for sampling tables with fixed marginals and weighted row sums. This strategy can be used to estimate exact significance levels for a number of conditional tests. It can also give an estimate of the total number of tables under these constraints. The technique is illustrated on several examples.

**Merlise Clyde**
Duke University
Institute of Statistics & Decision Sciences

"Stochastic Search via Sampling Adaptively without Replacement"

With the advent of Markov chain Monte Carlo algorithms, variable and model selection in high dimensional problems became feasible for a wide class of problems. Estimation of model probabilities via Monte Carlo frequencies, however, was often not satisfactory and suffered from slow convergence. Alternative estimates of model probabilities based on using exact marginals when available or numerical approximations normalized over the set of distinct models sampled proved to be superior for model selection and model averaging. Rather than using MCMC methods that may revisit previously sampled models, we propose sampling without replacement. Ideally, we would like to sample models proportional to their posterior probability. While this is initially unknown, we can construct adaptive proposals based on the previous history of the sample for generating new models. We describe several approaches for sampling adaptively without replacement (SAR), discuss their computational requirements, and how they can be used for model search and model averaging.

**Adrian Dobra**
Duke University
Institute of Statistics & Decision Sciences

Compositional Regressions, DAGS and Fitting High-Dimensional Gaussian Graphical Models

Finding graphical models for datasets with thousands of variables when the sample size is extremely small raises many theoretical and computational issues. In this talk I will identify some of these issues relating to model search as well as with displaying the results of the analyses. I will present a novel algorithm for finding graphical models that takes advantage of parallel computing. I will end with open questions for future research.

**Jean-Pierre Fouque**
North Carolina State University
Department of Mathematics

Multi-Scale Stochastic Volatility

In this talk I will present stochastic volatility models which incorporate several factors on different well-separated time scales. Historical returns data and options data show that two time scales are needed, one fast and one slow. I will then explain how one can deal with these two factors by using perturbation theory.

**Luis David Garcia**
Virginia Tech
Department of Mathematics

Algebraic Geometry of Bayesian Networks with Hidden Variables

The set of probability distributions that satisfy a conditional independence statement is the zero set of certain polynomials and can hence be studied using methods from algebraic geometry. We call such a set an independence variety. In this talk, I will focus on a particular probabilistic model based on directed acyclic graphs known as Bayesian Networks. I will describe the

polynomials defining these independence varieties and present some fundamental algebraic problems about them. Particular attention will be paid to models with hidden variables.

Joint work with Bernd Sturmfels and Mike Stillman

**Sujit K Ghosh**
North Carolina State University
Department of Statistics

Bayesian Analysis of Random Coefficient Autoregressive Models

A class of flexible Bayesian methods is presented to obtain parameter estimates and test non-stationarity (unit-root) hypothesis for Random Coefficient Autoregressive (RCA) models. RCA models are obtained by introducing random coefficients to an AR or more generally ARMA models. These models have second order properties similar to that of ARCH and GARCH models hence can be viewed as a volatility model. We investigate two model selection criteria and show that RCA models are robust against model misspecifications. We introduce a flexible class of priors for the stationarity parameter of RCA models and propose couple Bayesian unit-root testing procedures. Simulation results show that our proposed test procedures have good frequentist properties in terms of achieving high statistical power while maintaining low total error rates. Finally, a real life example involving exchange rates is presented to show the applicability of the proposed methods.

Joint work with Dazhe Wang

**Paolo Giudici**
University of Pavia
Department of Statistics

MCMC Methods for Bayesian Data Mining

In the talk I will give examples of complex data analysis which have been solved, under a Bayesian approach, by means of MCMC computations. These will concern, in particular, model selection and comparison for graphical Markov models. I will outline how reversible jump MCMC can be designed and developed for the previous problem in a rather efficient way. I will then consider ways to improve the efficiency of the MCMC approximation, especially in the presence of a large number of random variables. I will show how a recent approach, proposed in Brooks, Giudici and Roberts, 2003, can improve the performance of standard reversible jump approach. The paper will also contain a discussion on the choice of appropriate convergence diagnostic algorithm for reversible jump MCMC applications.

**Sean Han**
North Carolina State University
Department of Mathematics

Pricing Asian Options and Variance Swaps with Volatility Scales

In this talk, we first present a dimension reduction technique for pricing arithmetic average Asian options in the context of multiscale-factor stochastic volatility models. This technique is very useful to reduce computational efforts comparing to the usual higher dimensional Asian pricing problems. In particular, the price approximation derived from the multiscale asymptotic analysis allows relevant parameters to be calibrated from the implied volatility surface. We then turn to a similar subject: pricing volatility contracts but the SV model is restricted to one fast mean-reverting factor. We show that applying the Feynman-Kac formula may lead to a degenerated pricing PDE. A particular contract corridor is considered, and we show that the price of this contract should take account the local time appearing around boundaries of the corridor.

**Mark Huber**
Duke University
Department of Mathematics

Markov Chain Moves for Generating Contingency Tables with Fixed Weighted Row Sums

Monte Carlo algorithms for estimating exact p-values for various logistic regression models require generation of samples drawn from contingency tables subject to a constraint involving a weighted row sum. Previous attempts to construct Markov chains for this problem used far too many moves to be practical. Here we present two new approaches to this problem. The first uses a much simpler Markov chain, but at the price of the sample space being larger than our desired set. We show that this chain is irreducible, in some instances rapidly mixing, and that in practice the state space sampled from is not too much larger than the target space. The second approach uses a acceptance/rejection scheme from a larger space for which we can obtain samples using perfect sampling techniques.

**Chuanshu Ji**
University of North Carolina
Department of Statistics

Gaussian Approximations in MCMC Computation of Option Prices with Stochastic Volatility Models

A well-known challenging problem in empirical finance is to calibrate stochastic volatility (SV) models based on both return and option data. The major difficulty lies in the need for performing high-dimensional numerical integration for the option price in every iteration of a proposed algorithm --- efficient methods of moments (EMM) or Markov chain Monte Carlo (MCMC). Most works in the literature are only restricted to using the Heston's model that enjoys a closed-form solution. Alternatively, some attempt at other SV models using "brutal force" simulation encounters enormous computational intensity.

In this talk, we present an approximation scheme based on various central limit theorems (CLT) applied to certain additive functions of strong mixing sequences of the latent volatility under a risk-neutral probability measure. A great advantage of this method is to reduce the computation of option prices to very low-dimensional numerical integration. We will show some preliminary numerical results and indicate a much greater extent, including SV model calibration, option pricing, etc., to which the proposed approximation methodology can be applied.

**Beatrix Jones**
SAMSI

Adventures in Graph Space: Practical Issues in Exploring Graphical Structures for Moderate to Moderately Large Numbers of Variables

This talk will address practical issues in "scaling up" methods for evaluating graph marginal likelihoods and finding high posterior probability graphs. We will discuss appropriate priors on graph structures, issues in estimating marginal likelihoods for non-decomposable graphs, and contrast 2 algorithms for exploring graph space. Results from fitting graphs to two simulated data sets and a 150 variable gene expression data set will be presented.

**Feng Liang**
Duke University
Institute of Statistics & Decision Sciences

Mixtures of G-Priors and Variable Selection

Zellner's (1986) g-prior is widely used in variable selection in linear regression. An extension of the g-prior approach is to introduce a prior on the hyper-parameter g, for example, the Zellner-Siow (1980) prior corresponds to the Inverse-Gamma prior on g. In this talk, we will introduce a family of priors on g, called Hyper-G prior, because its marginal density can be expressed as the Hyper-Geometric function. It is easy to compute due to the close-form marginal and it has been shown to be consistent in model selection and Bayesian model averaging. Its corresponding methods in variable selection, null-model based and full-model based, will be compared with Zellner-Siow prior on some real data.

**Helene Massam**
York University

Graphical Gaussian Models Selection in a Bayesian Framework

We consider a given set of variables $X = (X_1, \ldots, X_n)$. We assume that $X$ follows a centered Gaussian model $N(0,\Sigma)$ and that the various dependences and independences between the variables $X_i, i = 1, \ldots, n$ can be represented by means of an undirected graph $G = (V,E)$ with $V = \{1, \ldots, n\}$ the set of vertices, each vertex representing a variable and with $E$ the set of edges, in the following way. The variables $X_i$ and $X_j$ are independent given $X_{V\setminus\{i,j\}}$ if and only if the edge $(i,j)$ is not in $E$, that is, $(\Sigma^{-1})_{ij} = 0$ if and only if the edge $(i,j) \notin E$. Let $M_G^+$ be the cone of $p \times p$ positive definite matrices with zero entries whenever $(i,j) \notin E$.

A graphical Gaussian model ${\cal M}_G$ is the model

$$M_G = \{N(0,\Sigma); \Sigma^{-1} \in M_G^+\}$$

Let G be the set of all possible undirected graphs with set of vertices $V$. Given a sample from the distribution of $X$, choosing a graphical Gaussian model is equivalent to choosing a graph $G$ in G.

We work in a Bayesian framework. The distributions a priori are a Wishart distribution on $M_G^+$ for $K = \Sigma^{-1}$, given $G$, with density proportional to

$$(\det K)^{\frac{\delta-2}{2}} \exp -\frac{1}{2}K,$$

and the uniform density on G for $G$. We present a Markov chain on G which has as its stationary distribution the posterior distribution on G. Knowing this posterior distribution will allow us to select the models with highest density, given the data.

**German Molina**
Duke University
Institute of Statistics & Decision Sciences

MCMC for Estimation of Multiscale Stochastic Volatility Models

Financial returns can be driven by several volatility processes defined at different time scales. We develop an algorithm to estimate stochastic volatility models with several volatility seriess and test it under different scenarios. We also compare it with the traditional simple stochastic volatility model. We show the different results obtained in the estimation of the mean-reverting and vol-vol parameters when a second volatility process is introduced.

**Tao Pang**
North Carolina State University
Department of Mathematics

A Stochastic Computational Method for Portfolio Optimization Problems

Some portfolio optimization problems can be formulated as stochastic control problems. The goal is then to find a solution of the derived dynamic programming equation (DPE). In this presentation, a stochastic computational method to solve the DPE is discussed. The key idea is to use Markov chain approximation. Some techniques to implement this method is also studied. The numerical results are given and results turn out to be very consistent with the theoretical results.

**Rui Paulo**
SAMSI

Marginal Likelihoods and Bayes Factors

We consider the problem of computing marginal likelihoods, particularly in the context of model selection. For that matter, we describe and compare several stochastic computation methods that aim at computing marginal likelihoods, ratios of normalizing constants or posterior model probabilities, having in mind that the final goal is to characterize the posterior distribution on the model space.

Such methods can be divided in two groups. One group requires a single Markov Chain and no model enumeration. The other group requires model enumeration since all marginal likelihoods (or ratios of normalizing constants) are computed one at a time.

After describing the methods, highlighting some of their deficiencies and advantages, we illustrate their use in the context of the probit regression model, using both real and simulated data.

**Sujit Sahu**
University of Southhampton
Department of Mathematics

"Stochastic Simulation Methods for the Number of Components in a Mixture"

The problem of determining the unknown number of components in mixtures is of considerable interest to researchers in many areas. This paper generalizes a Bayesian testing method based on the Kullback-Leibler distance proposed by Mengersen and Robert (1996). An alternative, weighted Kullback-Leibler distance is proposed as testing criterion. Explicit formulas for this distance are given for a number of mixture distributions. A step-wise testing procedure is proposed to select the minimum number of components adequate for the data. A fast, collapsing approach is proposed for reducing the number of components which does not require full re-fitting at each step. The method, using both distances, is compared to the Bayes factor approach. The method is easy to implement and is illustrated using BUGS, a general purpose software for Bayesian analysis.

**Aleksandra Slakovic**
Carnegie Mellon University
Department of Statistics

"Making Inferences from Arbitrary Sets of Conditionals and Marginals for Contingency Tables"

In recent work on statistical disclosure limitation and confidentiality, Dobra and Fienberg (2000,2002) have employed Groebner bases, in connection with decomposable and graphical log-linear models given a set of margins, to establish bounds and distributions for the cell entries in contingency tables. Building on their work, we explore connections between log-linear and graphical models such as DAGs and algebraic structures when the space of interest is defined by a collection of conditionals and marginals. We develop upper and lower bounds for cell entries given arbitrary sets of marginal and conditional distributions and describe moves (data swaps) that maintain these fixed sets of distributions. We give a complete characterization of the two-way table problem and discuss extensions to multi-way tables.

**Seth Sullivant**
University of California, Berkeley
Department of Mathematics

"Markov Bases of Binary Graph Models"

We explore Markov bases for multidimensional binary tables where are marginals computed are 2-way marginals. Such a computation of 2-way marginals is naturally encoded by a graph. A "classic" theorem says that there is a quadratic Markov basis if and only if the underlying graph is a forest. We investigate graphs with loops, and describe Markov bases for cycles and for the

bipartite graph K_2,n. Furthermore, we investigate the combinatorial structure of Markov basis elements of various degrees.

Joint work with Mike Develin

**Ruriko (Rudi) Yoshida**
University of California, Davis
Department of Mathematics

□Variations on Barvinok's Counting Algorithm and Applications to Multiway Contingency Tables□

In 1993 Barvinok gave an algorithm that counts lattice points in convex rational polyhedra in polynomial time when the dimension of the polytope is fixed.

In the first half of this talk we will present what we call the homogenized Barvinok's algorithm. As the original Barvinok's algorithm, it runs in polynomial time to the input size when the number of variables is fixed, but it has some computational advantages: (1) when dealing with the polytopes with few facets but large number of vertices, (2) in the computation of the Hilbert series of Ehrhart rings associated to rational polytopes. We will present the new results with some contingency tables via the homogenized Barvinok algorithm.

In 2002 Barvinok and Woods described how to efficiently obtain, for fixed dimension, the Hilbert basis of a simplicial cone.

In the second half of this talk, with Barvinok's and Woods' theorem, we will describe that in fixed dimension we can find Markov bases of contingency tables in polynomial time on the size of the input.

**Hao Helen Zhang**
North Carolina State University
Department of Statistics

□Nonparametric Regression Variable Selection □COSSO□

A new method "COSSO" will be introduced for model selection and model fitting in nonparametric regression models. Built in the framework of smoothing spline analysis of variance, the "COSSO" means COmponent Selection and Smoothing Operator. It is a method of regularization with the penalty functional being the sum of component norms, instead of the squared norm employed in the traditional smoothing spline method. The COSSO provides a unified framework for several recent proposals for model selection in linear models and smoothing spline analysis of variance models. Theoretical properties in terms of estimation and model selection are studied. The COSSO is compared with the MARS in simulations and real examples, and is shown to give very competitive performances.

## IV. DATA MINING AND MACHINE LEARNING

### *A. Opening Workshop Program*
September 6-10, 2003

**Saturday – September 6, 2003**
MCNC-RDI Auditorium

| | |
|---|---|
| **12:30-1:00 pm** | Registration Check-in |
| **1:00-2:30 pm** | *Tutorial 1:* Large p, Small n Inference<br>**David Banks**, Duke University |
| **2:30-3:00 pm** | Coffee Break |
| **3:00-4:30 pm** | *Tutorial 2:* Support Vector Machines<br>**J. S. Marron**, University of North Carolina-Chapel Hill |

**Sunday – September 7, 2003**
Radisson Governors Inn, Room H

| | |
|---|---|
| **8:30-9:30 am** | Registration Check-In |
| **9:30-10:00 am** | Welcome and Introductions<br>**Jim Berger**, Director of SAMSI<br>**Alan Karr**, Director of NISS & DMML Program Leader |
| **10:00-10:45 am** | Similarities and Differences between Statistics, Machine Learning and Data Mining<br>**Leo Breiman**, University of California-Berkeley |
| **10:45-11:00 am** | Coffee Break |
| **11:00-11:45 am** | Convex Optimization and Variational Inference Algorithms: Alternatives to MCMC Large-Scale Statistical Models<br>**Michael Jordan**, University of California-Berkeley |
| **12:00-1:00 pm** | Lunch |
| **1:00-3:00 pm** | *Birds-of-a-Feather Session* (Precursors of Working Groups)<br><br>Topics will reflect workshop and participant interests. Current possible topics include Large p, Small n Inference, Bioinformatics, Support Vector Machines, Computational Experiments, Text Mining and Model Selection. |
| **3:00-3:30 pm** | Coffee Break |

| | |
|---|---|
| **3:30-4:15 pm** | Statistical Methods for Text Mining<br>**David Madigan**, Rutgers University |
| **4:30-5:15 pm** | Using Proc MULTEST of SAS/STAT for Data Mining<br>**Peter Westfall**, Texas Tech University |

**Monday – September 8, 2003**
Radisson Governors Inn, Room H

| | |
|---|---|
| **8:30-9:30 am** | Registration Check-In |
| **9:30-10:15 am** | Temporal Data Mining: Novel Algorithms and Their Applications<br>**K. P. Unnikrishnan**, General Motors Research & Development |
| **10:30-11:00 am** | Coffee Break |
| **11:00-11:45 am** | Postmarketing Drug Adverse Event Surveillance and the Innocent Bystander Effect<br>**William DuMouchel**, AT&T Labs Research |
| **12:00-1:00 pm** | Lunch |
| **1:00-2:00 pm** | *Poster Sales Talks* (2 minutes each) |
| **2:00-2:30 pm** | Coffee Break |
| **2:30-3:15 pm** | Data Mining in Anti-Terrorism Applications<br>**Jeff Schneider**, Carnegie Mellon University |
| **3:30-4:00 pm** | Coffee Break |
| **4:00-5:30 pm** | *Second Chance Seminar* (Open Floor Session) |
| **7:00-9:00 pm** | *Reception & Poster Session* @ NISS-SAMSI Building<br>(19 TW Alexander Drive, 685-9350)<br>Transportation provided by Carolina Livery from 6:45-9:15pm |

**Tuesday – September 9, 2003**
Radisson Governors Inn, Room H

| | |
|---|---|
| **8:30-9:30 am** | Registration Check-In |
| **9:30-10:15 am** | Statistical Tools for the Sciences<br>**Leo Breiman**, University of California-Berkeley |

| | |
|---|---|
| **10:30-11:00 am** | Coffee Break |
| **11:00-11:45 am** | "Using Graphics in Exploratory Data Analysis and Data Mining: An Application of Supervised Classification in Olive Oil Quality"<br>**Di Cook**, Iowa State University |
| **12:00-1:00 pm** | Lunch |
| **1:00-3:00 pm** | *Young Researchers Session* |
| **3:00-3:30 pm** | Coffee Break |
| **3:30-4:15 pm** | "Bayesian Additive Regression Trees"<br>**Robert McCulloch**, University of Chicago |
| **4:30-5:00 pm** | *Final Discussion* |
| **5:00 pm** | Adjourn |

**Wednesday – September 10, 2003**
NISS-SAMSI Building, Room 104

| | |
|---|---|
| **9:30-10:00 am** | Arrival & Continental Breakfast |
| **10:00-12:00 pm** | Opening Remarks and Division into Groups |
| **12:00-1:00 pm** | Lunch |
| **1:00-3:00 pm** | Working Groups |

*B. Opening Workshop Abstracts*

**Leo Breiman**
University of California, Berkeley
Department of Statistics
*leo@stat.berkeley.edu*

"Statistical Tools for the Sciences"

With floods of data beating on our shores, there is an increasing need to provide scientists with effective tools to extract the relevant information from the relevant parts of this flood. I, and my coworker, Adele Cutler, have been engaged as toolmakers over the last

two-three years developing a suite of general purpose tools going under the generic name of "random forests", available as free open source.

I call these general purpose because they not only form accurate predictions, but a host of other functions valuable to the sicences such as computing variable importance and interactions, finding outliers, effectively replacing missing values, giving insightful projected views of the data, and unsupervised clustering. Thousands of copies have been downloaded and are being used in fields as diverse as analysis of microarray data and document classification.

**Di Cook**
Iowa State University
Department of Statistics
*dicook@iastate.edu*

Using Graphics in Exploratory Data Analysis and Data Mining: An Application of Supervised Classification in Olive Oil Quality

This talk is an introduction to interactive data visualization as it is practiced as part of exploratory data analysis and data mining. We discuss the use of graphical methods for supervised classification - brushing of simple scatterplots, rotations in higher dimensions such as the grand tour, directed searches in higher dimensions for interesting low dimensional views using projection pursuit, and manually controlled searches. We will describe how to use these methods in association with commonly used statistical methods such as linear discriminant analysis, trees and data mining techniques such as neural network and support vector machine classifiers.

A data set on olive oil quality will be used to coordinate the threads of the talk. There's an interesting story that we've learned about this data. GGobi, publicly available dynamic visualization software (http://www.ggobi.org), will be used.

Statistics and data mining allow us the fun of uncovering interesting and unexpected aspects of our world.

**William DuMouchel**
AT&T Labs--Research, Florham Park, NJ
*dumouchel@research.att.com*

Postmarketing Drug Adverse Event Surveillance and the Innocent Bystander Effect

The Multi-item Gamma Poisson Shrinker (MGPS) is an empirical Bayesian method for identifying unusually frequent counts in a large sparse frequency table. This presentation focuses on estimating associations among drugs and adverse event codes in databases of postmarketing reports of adverse drug reactions, as practiced by FDA and other safety

researchers. Extended methods can be used to signal frequent itemsets with more than two items, such as combinations of two drugs and one AE, or syndromes of multiple AEs. Another extension allows us to focus on detecting differences between itemset frequencies in different subsets of the data, or from one time period to another. Recent research attempts to adjust drug-adverse event associations for the effects of concomitant medications-sometimes called the "innocent bystander problem."

**Michael I. Jordan**
University of California, Berkeley
Division of Computer Science and Department of Statistics
*jordan@cs.berkeley.edu*

"Convex Optimization and Variational Inference Algorithms: Alternatives to MCMC for Large-Scale Statistical Models"

Tools from convex analysis and convex optimization are providing an increasingly important role in linking statistics and computer science. Convexity allows powerful computational tools to be brought to bear in solving large-scale statistical problems, but it also provides mathematical structure upon which theoretical analysis can be built. I will overview several examples of these links: (1) methods for probabilistic inference in large-scale graphical models based on convex relaxations, (2) algorithms for classification in problems involving multiple, heterogeneous data types based on semidefinite programming, and (3) theoretical analysis of the statistical consequences of basing classification algorithms on convex relaxations. I will illustrate these ideas with examples from information retrieval and bioinformatics.

[Joint work with Peter Bartlett, Nello Cristianini, Laurent El Ghaoui, Gert Lanckriet, Jon McAuliffe and Martin Wainwright]

**David Madigan**
Rutgers University
Department of Statistics
*madigan@stat.rutgers.edu*

"Statistical Methods for Text Mining"

Statistical methods for the analysis of textual data have a rich history yet rarely appear on standard Statistics curricula. Fueled by computing power, the last decade or so has seen some dramatic advances. This talk will examine some simple and not-so-simple statistical methods for part-of-speech tagging, information extraction, and text categorization. The text categorization application falls within the "high dimension, low sample size" class and presents some unique challenges.

**Robert E. McCulloch**
University of Chicago
Graduate School of Business
*robert.mcculloch@gsb.uchicago.edu*

We consider a Markov chain Monte Carlo (MCMC) algorithm for building an additive model with a sum of trees. The trees themselves are \treed models" [1], with a separate linear regression model in each terminal node. To adopt a Bayesian approach, we put prior distributions on parameters within each tree and on the sum of trees. The MCMC algorithm used in [1] to train a single tree is extended to the additive framework. The key component of this extension is a step in which a single random tree is drawn conditional on all others in the sum. The extension is straightforward yet powerful, enabling a more flexible set of models.

The model and associated training algorithm have some interesting similarities to Boosting and backfitting. If the priors are set so as to heavily regularize individual trees, we see Boosting-like behaviour with a large number of weak learners, each contributing a small amount to the overall model. Since a treed regression is anything but weak, careful attention must be paid to the choice of prior parameters. If instead we relax the regularization, then a smaller number of additive trees will contribute to the model. The iterated draws of each tree conditional on others is similar to a Bayesian version of backfitting [2].

The Bayesian framework and MCMC training algorithm yield a posterior distribution, which can be used to assess uncertainty. For example, posteriors for the number of weak learners and for predictions are easily available.

References

[1] Chipman, H., George, E. and McCulloch, R. (2002) \Bayesian Treed Models", Machine Learning, 48, 299-320.
[2] Hastie, T. , and Tibshirani, R. (2000), \Bayesian Backfltting (with comments and a rejoinder by the authors)", Statistical Science, 15 (3) , 196-223

Joint work with H. A. Chipman, University of Waterloo and E. I. George, University of Pennsylvania

**Jeff Schneider**
Carnegie Mellon University
Robotics Institute
*jeff.schneider@cs.cmu.edu*

□Data Mining in Anti-Terrorism Applications□

The AUTON lab at Carnegie Mellon University is doing data mining research in two separate anti-terrorism application areas and this talk will give an overview of both.

If a terrorist releases a biological agent in a major city, detecting that attack even a few hours earlier could save thousands of lives. The biosurveillance project is developing algorithms for online detection of an attack from aggregate data sources such as emergency department admissions, drug store purchases, and absenteeism rates. The idea is that the overall behavior of a population will have noticable changes much sooner than the default method of waiting until people are extremely ill or dead and having medical doctors determine the cause. The statistical challenge is to develop models of the typical behavior of thousands of inter-related attributes and find deviations from it as early as possible without excessive false alarms. I will describe approaches based on scan statistics, Bayesian networks, and rule learning systems.

The link detection problem requires one to identify underlying collaborative relationships from a large data base of interactions between people. Such interactions may be very informative (such as a phone call) or largely coincidental (such as traveling to the same city at the same time). I will describe a probabilistic generative model used to identify such relationships and show how this identification may be used in a larger terrorist plot detection problem. Finally, I will look at expanding these ideas to a Bayesian network based method of identifying terrorist activity.

**K.P. Unnikrishnan**
General Motors Research & Development Center
*unnikri@hotmail.com*

Temporal Data Mining: Novel Algorithms and Their Applications

Discovering sequential and temporal patterns in data is an important task. This is usually done by discovering the "frequent episodes" in sequences of events. These "episodes" characterize interesting collections of events occurring relatively close to each other in some partial order. Current methods to do the above use only information about the ordering of events; the events themselves are assumed to be instantaneous. But in many applications (for example, time-stamped status logs of a manufacturing process), the events have durations and the event durations and inter-event intervals carry useful information. In this talk, we present a framework
for mining such data and describe its use in analyzing a manufacturing system.

**Peter H. Westfall**, John and Marguerite Nivers and Paul Whitfield Horn Professor of Statistics, Jerry S. Rawls College of Business Administration
Texas Tech University
Department of Information Systems and Quantitative Sciences
*westfall@ba.ttu.edu*

Using PROC MULTTEST of SAS/STAT(R) for Data Mining

PROC MULTTEST of SAS/STAT(R) is a useful tool for data mining (DM) applications where the goal is to identify discoveries, but also provide some degree of error control. The scope of DM applications to which it may be applied is broad: it has been used for neuro-imaging, genomics (as described in the latest Statistical Science, for example), multiple adverse events in patient subgroups, and animal carcinogenicity. The theory, methods and output of the software are described with particular emphasis on DM applications and control of false discoveries.

C. *Data Mining Workday on Support Vector Machines Program*
NISS-SAMSI Building
January 28, 2004

| | |
|---|---|
| **8:30-9:00 am** | Continental Breakfast |
| **9:00-9:05 am** | Welcome to SAMSI<br>**Alan Karr** and **David Banks** |
| **9:05-10:05 am** | Invited Speaker: **Ji Zhu,** University of Michigan |
| **10:05-10:15 am** | Coffee Break |
| **10:15-10:45 am** | Bayesian SVM Cluster Presentation<br>Speaker: **Ernest Fokoue**, SAMSI |
| **10:45-11:00 am** | Discussion<br>Leader: **Ernest Fokoue**, SAMSI |
| **11:00-11:15 am** | Coffee Break |
| **11:15-11:45 am** | Kernel Selection Cluster Presentation<br>Speaker: **Helen Zhang**, North Carolina State University |
| **11:45-12:00 pm** | Discussion<br>Leader: **Marc Genton**, North Carolina State University |
| **12:00-1:00 pm** | Lunch |
| **1:00-1:30 pm** | Feature Selection Cluster Presentation<br>Speaker: **Xiaodong Lin**, SAMSI |
| **1:30-1:45 pm** | Discussion<br>Leader: **Jeongyoun Ahn**, University of North Carolina |
| **1:45-2:00 pm** | Coffee Break |

| | |
|---|---|
| **2:00-2:30 pm** | Space-Time Cluster Presentation<br>Speaker: **Peng Liu**, North Carolina State University |
| **2:30-2:45 pm** | Discussion<br>Leader: **Marc Genton**, North Carolina State University |
| **2:45-3:00 pm** | Coffee Break |
| **3:00-4:00 pm** | Invited Speaker: **Tong Zhang**, IBM |
| **4:00-4:15 pm** | Coffee Break |
| **4:15-5:00 pm** | Panel Discussion<br>SVM: Open Problems and Future Directions<br>Panelists: **Atina Brooks, Helen Zhang, Tong Zhang** and **Ji Zhu** |

*D. Data Mining Workday on Support Vector Machines Abstracts*

**Ernest Fokoue**
SAMSI

Bayesian Analysis of SVM and SVM-like techniques: The Present and Some Ideas for the Future

In the first part of my presentation, I intend to provide an overview of existing techniques and tools, along with the main results and challenges, and possibly some elements of comparison of performance. The second part of my talk will be mainly speculative, with suggestions of possible extensions and improvements of existing techniques, along with many interesting questions that seem to naturally arise with this class of machine learning techniques.

**Xiaodong Lin**
SAMSI

Variable Selection for SVM using SCAD penalty.

We suggest a new regularization method for variable selection in SVM. The idea is to replace the lasso-type $L_1$ penalty by a nonconcave penalty called SCAD (smoothly clipped absolute deviation), proposed by Fan and Li (2002). The objective function can be locally approximated by a quadratic function and the minimization problem can be solved iteratively. We compare the performance of SCAD SVM with the standard linear SVM and some linear SVM methods with feature selection.

**Peng Liu**
North Carolina State University

"Mining Space-time Data"

This project is conducting experiments on a space-time dataset provided by NCAR. The purpose is to forecast ceiling and visibility over hours. The experiments we have done are based on classification and regression trees (CART) and support vector machines (SVM). The data set is relatively huge, with a lot of missing values. Moreover, the aspects of time and space dependence also challenge the methods like CART and SVM. This talk will present our current results and generate discussions for future work.

**Helen Zhang**
North Carolina State University

"Compactly supported kernels"

There are many choices of kernels to implement nonlinear SVMs. We consider compactly supported RBF kernels, which has a potential for computation improvement in kernel-based learning methods. For any kernel $K$, a series of compactly supported kernels $K_C$ can be derived by a certain thresholding on entries ($C$ is the threshold value). The choice of the threshold is very important, since it balances the trade-off between the sparsity associated with $K_C$ and the similarity between $K_C$ and $K$. This work aims to develop some practical criteria to choose an optimal $C$. In the context of support vector machines, the performance of compactly supported kernels is shown in some simulation examples, and compared with that of the regular RBF kernels.

**Tong Zhang**
IBM

"Statistical Models for Binary and Multi-category Large Margin Methods"

Large margin classification methods, such as support vector machines, have been successfully applied to many pattern recognition problems. Practical implementations of these methods often lead to empirical minimizations of certain convex loss functions. In this talk, I discuss the following aspects of convex risk minimization based binary and multi-category classification methods:

1. Infinity-sample theory: statistical models

   We show that a classification scheme that minimizes a convex risk function induces a conditional probability estimate. Historically, the idea of margin maximization was motivated by considering completely separable classification problems. Consequently, unlike maximum likelihood estimate, some large margin methods has the ability to model zero-density directly without causing

robustness problems. We compare statistical models of various risk minimization formulations, and discuss their implications.

2. Finite-sample theory: consistency and rate of convergence

We show that by approximately minimizing a risk function which satisfies mild conditions, we also approximately minimize the classification error. Moreover, bounds can be established between the classification error of a classifier and its corresponding convex risk. Using such results, universal consistency and rate of convergence can be obtained for risk minimization based classification methods.

**Ji Zhu**
University of Michigan

□Piecewise linear SVM paths□

The support vector machine is a widely used tool for classification. In this talk, we consider two types of the support vector machine: the 1-norm SVM and the standard 2-norm SVM. Both types can be written as regularized optimization problems. In all current implementations for fitting an SVM model, the user has to supply a value for the regularization parameter. To select an appropriate regularization parameter, in practice, people usually pre-specify a finite set of values for the regularization parameter that covers a wide range, then either use a separate validation data set or use cross-validation to select a value for the regularization parameter that gives the best performance among the given set. In this talk, we argue that the choice of the regularization parameter can be critical. We also argue that the 1-norm SVM may have some advantage over the standard 2-norm SVM under certain situations, especially when there are redundant noise features. We show that the solution paths for both the 1-norm SVM and the 2-norm SVM are piecewise linear functions of the regularization parameter. We then derive two algorithms, respectively for the 1-norm SVM and the 2-norm SVM, that can fit the entire paths of SVM solutions for every value of the regularization parameter, hence facilitate adaptive selection of the regularization parameter for SVM models.
It turns out that the piecewise linear solution path property is not unique to the SVM models. We will propose some general conditions on the generic regularized optimization problem for the solution path to be piecewise linear, which suggest some new useful predictive statistical modeling tools.

This is joint work with Saharon Rosset (IBM T.J.Watson), Trevor Hastie (Stanford U.), and Rob Tibshirani (Stanford U.)

*E. Data Mining Workday on Theory and Methods & Large p, Small n Inference Program*
NISS-SAMSI Building
February 4, 2004

| | |
|---|---|
| **8:30-9:00 am** | Continental Breakfast |
| **9:00-9:30 am** | Introduction to SAMSI and the Data Mining Year |
| **9:30-10:00 am** | Large p Small n<br>**Bertrand Clarke,** University of British Columbia, Duke University & SAMSI |
| **10:30-10:45 am** | Coffee Break |
| **10:45-11:15 am** | Multiple Testing<br>**David Banks**, Duke University |
| **11:15-11:45 am** | GM Data<br>**Ashish Sanil**, NISS<br>**Jen-hwa Chu**, Duke University<br>**Alan Karr**, NISS |
| **11:45-1:15 am** | Lunch |
| **1:15-2:00 am** | Invited Talk: **Giles Hooker**, Stanford University |
| **2:00-2:45 pm** | Invited Talk: **Hugh Chipman**, University of Waterloo |
| **2:45-3:00 pm** | Coffee Break |
| **3:00-3:45 pm** | Invited Talk: **Jim Cox**, SAS |
| **3:45-4:30 pm** | Invited Talk: **Regina Liu**, Rutgers University |
| **4:30-5:00 pm** | Summary |

*F. Data Mining Workday on Theory and Methods & Large p, Small n Inference Abstracts*

**Regina Liu**
Rutgers University

Mining Massive Text Data and Developing Tracking Statistics

This talk outlines a systematic data mining procedure for exploring large free-style text datasets to discover useful features and develop tracking statistics, generally referred to as performance measures or risk indicators. The procedure includes text mining, risk analysis, classification for error measurements and nonparametric multivariate analysis. Two aviation safety report repositories *PTRS* from the FAA and *AAS* from the NTSB will

be used to illustrate applications of our research to aviation risk management and general decision-support systems. Some specific text analysis methodologies and tracking statistics will be discussed.

*G. Data Mining Workday on Bioinformatics Program*
NISS-SAMSI Building
February 11, 2004

| | |
|---|---|
| 8:30-9:00 am | Continental Breakfast |
| 9:00-9:05 am | Welcome to SAMSI<br>**Alan Karr**, NISS |
| 9:05-9:20 am | Bioinformatics Year, Virtual Screening<br>**Stan Young**, NISS |
| 9:20-10:20 am | **Yvonne Martin,** Abbott Laboratories |
| 10:20-10:35 am | Discussion |
| 10:35-10:50 am | Coffee Break |
| 10:50-11:20 am | **Alex Tropsha**, University of North Carolina |
| 11:20-11:30 am | Discussion |
| 11:30-12:00 am | New Method for Parmacophone Mapping<br>**Jun Feng**, NISS |
| 12:00-12:10 pm | Discussion |
| 12:10-1:10 pm | Lunch |
| 1:10-2:10 pm | Towards Interpretability of Classifiers for Virtual Screening<br>**Will Welch**, University of British Columbia |
| 2:10-2:25 pm | Discussion |
| 2:25-2:55 pm | Structure/activity analysis of chemosensitivity variation based on bond arrangements<br>**Kerby Shedden**, University of Michigan |
| 2:55-3:05 pm | Discussion |

| | |
|---|---|
| 3:05-3:20 pm | Coffee Break |
| 3:20-3:40 pm | SVM applied to MAO dataset<br>**Atina Brooks**, North Carolina State University<br>**Scott Oloff**, University of North Carolina |
| 3:40-3:50 pm | Discussion |
| 3:50-4:35 pm | Applications of virtual screening as used by<br>• Hereditary Disease Foundation □**Jun Feng**, NISS<br>• LDDN □**Ke Zhang**, North Carolina State University |
| 4:35-4:45 pm | Discussion |
| 4:45-5:15 pm | Directed discussion and summary of the day□s events<br>**Yvonne Martin,** Abbott Laboratories<br>**Will Welsh,** University of British Columbia<br>**Alex Tropsha,** University of North Carolina<br>**Jackie Hughes-Oliver,** North Carolina State University |

*H. Data Mining Workday on Bioinformatics Abstracts*

**Kerby Shedden**
University of Michigan

□Structure/activity analysis of chemosensitivity variation based on bond arrangements□

Growth inhibition (GI) measurements of compounds against a diverse panel of cell lines provide valuable information about cell-type specific chemosensitivity, which is a highly relevant property for anti-cancer agents. We consider the problem of identifying mutual associations between chemical structure and GI profiles in a large database of compounds. A key difficulty is that only a very few biologically important GI profiles are available a priori, and so in general both the GI profile and the predictive substructure must be discovered in the data. We propose a method for discovering such relationships based on small "bond arrangements" which are only weakly informative alone, but can in some cases be sequentially expanded into substructures with high specificity for a particular GI profile. As a positive control, a familiar GI profile related to multidrug resistance is found to be associated with a tetravalent nitrogen motif. Other potentially interesting but harder to explain associations are found as well.

**Will Welch**
University of British Columbia

"Towards Interpretability of Classifiers for Virtual Screening"

Empirically, models which often perform well in terms of prediction in virtual screening make few assumptions about the functional form of the structure-activity relationship. These methods include K-nearest neighbours (KNN) and classification and regression trees (CART).

Such flexible models are, however, difficult to interpret in the following sense. In the presence of multiple activity mechanisms, we would like to know the important variables and their critical ranges for each mechanism.

Two methods will be described which improve interpretability in the above sense. For classification trees, a method of rearranging nodes removes redundant constraints (variables or limits) from the tree. The second method is a model averaging strategy, which averages over subsets of variables and identifies important subsets.

Joint work with Hugh Chipman, Marcia Wang, and Yan Yuan

## V. NETWORK MODELING FOR THE INTERNET

### *A. Workshops on Internet Tomography and Sensor Networks Program* (also funded by National Security Agency)
October 12-15, 2003

**Sunday – October 12, 2003**
Radisson Hotel Research Triangle Park

| | |
|---|---|
| **12:30-1:00 pm** | Registration and Check-in |
| **1:00-4:30 pm** | *Tutorial:* "Introduction to Internet Tomography"<br>**Rui Castro**, Rice University<br>**George Michailidis**, University of Michigan<br>**Matt Roughan**, AT&T |

**Monday – October 13, 2003**
Radisson Hotel Research Triangle Park

| | |
|---|---|
| **8:30-9:00 am** | Registration and Check-in |
| **9:00-10:30 am** | 5-Minute Madness Presentations<br>Chair: **Jim Berger**, SAMSI<br>*A series of 5-minute presentations by many participants on current research.* |

- Patrice Abry, *CNRS*
- Richard Baraniuk, *Rice University*
- Andre Broido, *CAIDA*
- Mark Coates, *McGill University*
- Mark Crovella, *Boston University*
- Rene Cruz, *University of California-San Diego*
- Alberto Grunbaum, *University of California-Berkeley*
- Al Hero, *University of Michigan*
- Alan Karr, *NISS*
- Eric Kolaczyk, *Boston University*
- Steve Marron, *SAMSI & UNC*
- Jean Meloche, *Avaya Labs*
- Vijay Nair, *University of Michigan*
- Matt Roughan, *AT&T Research*
- Jiayang Sun, *Case Western Reserve University*
- Cathy Xia, *IBM Research*
- Bin Yu, *UCLA*
- Linda Zhao, *University of Pennsylvania*

**10:30-11:00 am** Coffee Break

**11:00-12:00 pm** Theme Problem Presentation and Discussion Session I:
*Validation of Internet Tomography*
Chair: **J.S. Marron**, SAMSI

*11:00-11:30* Overview of the Avaya testbed deployed at UNC.
**Lorraine Denby** and **Jean Meloche**, Avaya

*11:30-12:00* Overview of network tomography, and inference about variable bandwidth
**Rob Nowak**, Rice University & University of Wisconsin

**12:00-1:30 pm** Lunch

**1:30-3:00 pm** Break Out Group Session I: *Prioritization of list of methods for validation*
Chair: **J.S. Marron**, SAMSI

**3:00-3:30 pm** Report of Breakout Groups
Chair: **J.S. Marron**, SAMSI

**3:30-4:00 pm** Coffee Break

**4:00-5:00 pm** What I wish I could do Session
Chair: **Alan Karr**, NISS & SAMSI

very short presentations, ideally discussion of problems perceived, but no ideas for solution appear to be available

- Mark Coates, McGill University
- Mark Crovella, Boston University
- Nick Duffield, AT&T Research
- Al Hero, University of Michigan
- Steve Marron, SAMSI & UNC
- Jean Meloche, Avaya Labs
- Jiayang Sun, Case Western Reserve University
- Don Towsley, University of Massachusetts
- Walter Willinger, AT&T Research

**Tuesday – October 14, 2003**
Radisson Hotel Research Triangle Park

**8:30-9:00 am** Registration and Check-in

**9:00-10:00 am** Theme Problem Presentation and Discussion Session II
Chair: **J.S. Marron**, SAMSI

- "Spatio-Temporal Network Data Collection and Analysis"
  **Christos Papadopoulos**, University of Southern California

- "COSSACK - Coordinated Suppression of Simultaneous Attacks"
  **Paul Barford**, University of Wisconsin

- "Global Characteristics and Prevalence of Internet Intrusion"

  *9:00-9:40* Overview
  *9:40-9:50* Discussion: **Bin Yu**, University of California-Berkeley
  *9:50-10:00* Discussion: **Walter Willinger**, AT&T

**10:00-10:30 am** Coffee Break

**10:30-11:30 am** Breakout Group Session II: "*Open Problems in Spatio-Temporal Network Data Collection and Analysis*"
Chair: **J.S. Marron**, SAMSI

**11:30-12:00 pm** Report of Breakout Groups
Chair: **J.S. Marron**, SAMSI

**12:00-1:00 pm** Adjournment of the Internet Tomography Workshop and Lunch

**1:00 pm** Opening of Workshop on Sensor Networks, October 14-15.

**1:00-2:30 pm** Current Research Presentations
Chair: **Alan Karr**, NISS & SAMSI
a series of 5 minute presentations by most participants, on their current work. Why are you here? Interests and on-going work.

- Richard Baraniuk, *Rice University*
- Jim Berger, *SAMSI & Duke University*
- Mujdat Cetin, *MIT*
- Mark Coates, *McGill University*
- Kendall Giles, *Johns Hopkins University*
- Alberto Grunbaum, *University of California-Berkeley*
- Al Hero, *University of Michigan*
- Thomas Lee, *Colorado State University*
- Benyuan Liu, *City College of New York*
- Steve Marron, *SAMSI & UNC*
- Jean Meloche, *Avaya Labs*
- Mohammad Rahimi, *UCLA*
- Srikant, *University of Illinois at Urbana-Champaign*
- Murad Taqqu, *Boston University*
- Don Towsley, *University of Massachusetts*
- Yolanda Tsang, *Rice University*
- Frank Vernon, *University of California-San Diego*
- Walter Willinger, *AT&T Research*

**2:30-3:00 pm** Coffee Break

**3:00-3:30 pm** Introduction to the applications and systems issues of Sensor Networks
**Deborah Estrin**, University of California-Los Angeles
Chair: **J.S. Marron**, SAMSI

**3:30-4:00 pm** Overview of statistical processing techniques for sensor networks
**Feng Zhao**, Palo Alto Research Center
Chair: **J.S. Marron**, SAMSI

**4:00-4:30 pm** Sensor Fusion
**Jose Moura**, Carnegie Mellon University
Chair: **Murad Taqqu**, SAMSI

**4:30-4:45 pm** Discussion
**John Fisher**, Massachusetts Institute of Technology

**7:00-9:00 pm** **Reception and Poster Session** for both workshops at

NISS-SAMSI (19 T.W. Alexander Drive, 919-685-9350)
Transportation from the Radisson to NISS-SAMSI will be provided by Carolina Livery from 6:45 until 9:15 pm.

**Wednesday – October 15, 2003**
Radisson Hotel Research Triangle Park

**8:30-9:00 am** Registration and Check-in

**9:00-9:30 am** "Multiple Target Tracking and Detection"
- **Kung Yao**, University of California-Los Angeles

Chair: **David Rolls**, SAMSI

**9:30-9:45 am** Discussion
- **Robert Wolpert**, Duke University

**9:45-10:00 am** Coffee Break

**10:00-10:30 am** Node Localization
- **Mani Srivastava**, University of California-Los Angeles

Chair: **George Michailidis**, University of Michigan

**10:30-10:45 am** Discussion
- **Al Hero**, University of Michigan

**11:00-12:00 pm** "Pie in the Sky session" aka "What I wish I could do" Session
Chair: **Jim Berger**, SAMSI
5-minute presentations, ideally discussion of problems perceived, but no ideas for solution appear to be available

- Al Hero, University of Michigan
- Steve Marron, SAMSI & UNC
- Mani Srivastava & Deborah Estrin, UCLA
- Frank Vernon, University of California-San Diego
- Cliff Wang, Army Research Office
- Kun Yao, UCLA
- Feng Zhao, Palo Alto Research Center

**12:00-1:00 pm** Lunch

**1:00-1:30 pm** Field Estimation

- **Rob Nowak**, Rice University & University of Wisconsin

Chair: **Cheolwoo Park**, SAMSI

**1:30-1:45 pm** Discussion

- **Mark Hansen**, University of California-Los Angeles

**1:45-2:00 pm** Coffee Break

**2:00-2:30 pm** Distributed Coding (compression)

- **Kannan Ramchandran**, University of California-Berkeley

Chair: **Krishanu Maulik**, Eurandom

**2:30-2:45 pm** Discussion

- **Andrew Nobel**, University of North Carolina

**3:00-3:30 pm** Efficient Design of Wireless Sensing Systems for Detection Applications

- **Venugopal Veeravalli**, University of Illinois & National Science Foundation

Chair: **Kevin Jeffay**, University of North Carolina

**3:30-3:45 pm** Discussion

- **Don Towsley**, University of Massachusetts

**4:00-4:30 pm** Summary: Where we can go from here?
**Rob Nowak**, Rice University & University of Wisconsin
Chair: **J.S. Marron**, SAMSI

**4:30 pm** Adjournment

*B. Workshop on Congestion Control and Heavy Traffic Modeling Program*
October 31-November 1, 2003

**Friday - October 31, 2003**
Radisson Hotel, Room H □3rd Floor

**8:30-9:00 am** Registration and Check In

**9:00-12:00 pm** **Tutorials**

*9:00-10:00* □Overview of SAMSI Measurement and Modeling Activies□
**J. S. Marron**, SAMSI

Chair: **Cheolwoo Park**, SAMSI

*10:00-10:15* Coffee Break

*10:15-11:45* "Congestion Control and Heavy Traffic Modeling"
**Ruth Williams**, University of California-San Diego
Chair: **David Rolls**, SAMSI

**11:45-1:00 pm** Lunch

**1:00-3:00 pm** 5-Minute Madness Presentations
Chair: **Jim Berger**, SAMSI

**3:00-3:30 pm** Coffee Break

**3:30-5:00 pm** **Theme Problem Presentations I**
Chair: **J. S. Marron**, SAMSI

*3:30-4:00* **Francois Baccelli**, ENS

*4:15-4:45* **Jim Dai**, Georgia Tech University

**5:00-5:15 pm** Discussion
**Armand Makowski**, University of Maryland

**7:00-9:00 pm** **Reception and Poster Session at SAMSI**
*(19 T.W. Alexander Dr., 919-685-9350)*
*Transportation from the Radisson to SAMSI will be provided by Carolina Livery from 6:45-9:15 pm. Shuttle will be located in front of the hotel either to the left or the right of the awning.*

**Saturday – November 1, 2003**
Radisson Hotel, Room BC "2nd Floor

**8:15-8:45 am** Registration and Check In

**8:45-10:30 am** **Theme Problem Presentations II**
Chair: **J. S. Marron,** SAMSI

*8:45-9:15* **Vishal Misra**, Columbia University

*9:30-10:00* **Kevin Jeffay**, University of North Carolina-Chapel Hill

*10:15-10:30* Discussion - Overview
**Don Towsley**, University of Massachusetts

**10:30-11:00 am** Coffee Break

**11:00-12:00 pm** Break Out Group Discussion: *Formulate and Prioritize Open Problems*
Chair: **J. S. Marron**, SAMSI

**12:00-1:30 pm** Lunch

**1:30-2:15 pm** Report of Breakout Groups
Chair: **J. S. Marron**, SAMSI

**2:15-3:00 pm** Pie in the Sky Session
Chair: **Jim Berger**, SAMSI

**3:00-3:30 pm** Coffee Break

**3:30-4:30 pm** New Direction Overview Talks
Chair: **Jim Berger**, SAMSI

- **Francois Baccelli**, ENS
- **Rolf Riedi**, Rice University
- **Don Smith**, University of North Carolina-Chapel Hill

**4:30 pm** Adjournment

## VI. MULTISCALE MODEL DEVELOPMENT AND CONTROL DESIGN

### *A. Opening Workshop Program*
January 17-20, 2004

**Saturday – January 17, 2004**
MCNC-RDI Auditorium

**11:00-12:00 pm** Registration and Check-in

**12:00-1:30 pm** *Tutorial:* Energy Techniques for Multiscale Modeling
**Ralph Smith**, North Carolina State University

**1:30-2:00 pm** Coffee Break

**2:00-3:30 pm** *Tutorial:* Principles of Multilevel Stochastic Modeling
**Alan Gelfand,** Duke University

**3:30-4:00 pm** Coffee Break

**4:00-5:30 pm** *Tutorial:* Control Design for Nonlinear Systems
**Art Krener,** University of California-Davis

**Sunday – January 18, 2004**
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

**8:00-8:30 am** Registration and Check-in

**8:30-10:30 am** **SESSION 1:** Modeling Issues in Multiscale Problems
Chair: **Ioannis Kevrekidis**, Princeton University

- **Don Brenner**, North Carolina State University
- **Hideo Mabuchi**, California Institute of Technology

**10:30-11:00 am** Coffee Break

**11:00-12:00 pm** **Poster Talks** (3 minutes each)

**12:00-1:00 pm** Lunch, Room A-B (2nd floor)

**1:00-2:00 pm** **Poster Talks** (continued)

**2:00-2:30 pm** Coffee Break

**2:30-4:30 pm** **SESSION 2:** Control Design in the Presence of Uncertainty
Chair: **Kirsten Morris**, University of Waterloo

- **Murti Salapaka,** Iowa State University
- **C.F. Jeff Wu,** Georgia Institute of Technology

**6:30-8:30 pm** **Reception and Poster Session** at NISS-SAMSI
(19 T.W. Alexander Drive, 919-685-9350)

***Poster Presenters****:* The Radisson Shuttle will depart at 5:45pm to bring to SAMSI to set-up your poster.

***All other Participants****:* Carolina Livery will be providing continuous shuttle service between the Radisson and SAMSI. The first shuttle will depart from the Radisson at 6:20pm and the last shuttle will leave SAMSI at 8:35pm.

**Monday – January 19, 2004**
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

**8:15-8:30 am** Registration and Check-in

**8:30-10:30 am** **SESSION 3:** Scaling Issues in Spatial Modeling
Chair: **Christopher Wikle**, University of Missouri

- **Montserrat Fuentes,** North Carolina State University
- **Carol Gotway Crawford,** Centers for Disease Control

**10:30-11:00 am** Coffee Break

**11:00-11:45 am** **SESSION 4:** Homogenization and Multiscale Modeling Issues
Chair: **Belinda King**, Oregon State University

- **Christopher Lynch,** Georgia Institute of Technology

**11:45-1:00 pm** Lunch, Room A-B (2nd Floor)

**1:00-2:15 pm** **SESSION 4 –** continued
Chair: **Belinda King**, Oregon State University

- **Doina Cioranescu,** CNRS-University of Paris 6

**2:15-2:45 pm** Coffee Break

**3:00-5:00 pm** **SESSION 5:** Multiresolution Function Estimation
Chair: **Brani Vidakovic,** Georgia Institute of Technology

- **Marina Vannucci,** Texas A&M University
- **Yazhen Wang,** University of Connecticut

**Tuesday – January 20, 2004**
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

**9:00-9:45 am** **SESSION 6:** Numerical Techniques for Multiscale Materials
Chair: **David Higdon,** Los Alamos National Laboratory

- **Chuanshu Ji,** University of North Carolina-Chapel Hill

| | |
|---|---|
| **9:45-10:15 am** | Coffee Break |
| **10:15-11:30 am** | **SESSION 6 –** continued<br>Chair: **David Higdon,** Los Alamos National Laboratory<br>• **Tom Hou,** California Institute of Technology |
| **11:30-12:00 am** | **Discussion** |
| **12:00-1:00 pm** | Lunch, Room A-B (2nd Floor) |
| **1:00-3:00 pm** | **Working Groups** |
| **3:00-3:30 pm** | Coffee Break |
| **3:30-5:00 pm** | **Working Groups** |

*B. Opening Workshop Abstracts*

**Donald W. Brenner**
North Carolina State University
Department of Materials Science and Engineering
*default@mindspring.com*

Multi-Scale Simulation Schemes in Materials Modeling*

Bridging disparate length and time scales has become a high-profile goal of the materials modeling community. Presented in this talk will be several concurrent and serial simulation schemes that are being developed to bridge these scales, as well examples of phenomena in materials science that scale naturally. Specific applications from the fields of microstructure modeling, nanofluidics, tribology, and indentation will be highlighted.

*This work is funded by the National Science Foundation through a Nanoscale Interdisciplinary Research Team and by the Office of Naval Research

**Doina Cioranescu**
CNRS-University Paris 6
Laboratoire J. L. Lions
*cioran@ann.jussieu.fr*

Homogenization of Multi-Scale Domains by the Periodic Unfolding Method

The aim of homogenization theory is to describe the behavior of composite materials that are commonly (and more and more) used in industry. The interest for these materials

comes from the fact that a composite has, in general, better characteristics than those of its components (typical examples are the reinforced concrete, the optical fibers or the superconducting multifilamentary materials). Computational methods, in the case where the composite has a very large number of heterogeneities, are difficult to implement and the discontinuities (or oscillations) constitute a source of errors. This is why one tries to describe the overall behavior of composite materials by taking into account the local characteristics of the heterogeneities, and the theory of the homogenization was precisely introduced to answer this challenge.

One can describe a composite at a local or ”microscopic” scale by taking into account the properties of each component. In the case where the composite has a very large number of heterogeneities, such a description, from numerical point of view for instance, becomes difficult and the discontinuities (or oscillations) constitute a source of errors. This is why one tries to describe the overall behavior of composite materials by taking into account the local characteristics of the heterogeneities, and the theory of the homogenization was precisely introduced to answer this challenge.

From mathematical point of view, one has to study the asymptotic behavior as $\varepsilon $ goes to zero, of partial differential equations whose coefficients are highly oscillating with respect to $\epsilon$. The parameter $\epsilon$ describes the heterogeneities, that are small compared to the global size of the composite. Another situation is that of perforated domains where it is the geometry which depends on $\epsilon$. The model case here is the periodic one: the heterogeneities (or the perforations) are periodically distributed with the period $\epsilon$.

Convergence and corrector results, error estimates as well as homogenization results in the framework of the periodic unfolding method will be presented. We will treat in particular the following examples: truss-like structures, the Stokes problem in porous media, multi-scale domains and the homogenization of non linear integrals.

**Carol A. Gotway Crawford**
National Center for Environmental Health
Centers for Disease Control and Prevention

Combining Incompatible Spatial Data: An Introductory Overview of Statistical Issues and Methods

The widespread availability of digital spatial data and the capabilities of geographic information systems make it possible to easily synthesize spatial data from a variety of sources. More often than not, data have been collected at different scales, and each of the scales may be different from the one of interest. Geographic information systems effortlessly handle these types of problems through raster and geoprocessing operations based on proportional allocation and centroid smoothing techniques. However, there are many statistical issues associated with combining such disparate data. This presentation provides an introductory overview of several such issues, including the problems that can

occur when making inferences from aggregated data, the concept of spatial support, and the importance of proper uncertainty assessment. From this perspective, the utility of many different statistical approaches to the problem of combining incompatible spatial data will be assessed.

Joint work with Linda J. Young, Department of Statistics, University of Florida

**Montserrat Fuentes** and Lian Xie, Professor of Oceanography, North Carolina State University
North Carolina State University
Department of Statistics
*fuentes@stat.ncsu.edu*

Data Assimilation Methods Using Disparate Spatial Information

Even when the fastest computers are used, atmospheric and oceanic numerical models are only approximations of the full dynamic and thermodynamic equations. Additionally, observations of atmospheric and oceanic variables are always disparate in space and time, so true initial or boundary conditions are never available. One way to improve the numerical model forecast is to apply a procedure known as data assimilation (DA), which blends the model approximations with observations in a least-square error sense. The standard methods for DA, e.g. Kalman Filter (KF), do not take into account the potential different spatial resolution of the model output and the sources of data, e.g., high-resolution radar data and low resolution model grids. Another limitation of the classical KF method is that assumes linearity of the numerical model and a Gaussian distribution.

We develop here formal DA methods to combine spatial disparate data with numerical model output. We present a new filter approach, that allows for nonlinearities and lack of normality and takes into account the multi-scale problem. This new method is be applied to an ocean model to forecast the response of the coastal ocean to hurricanes, by assimilating both high-resolution remote sensing data and low-resolution point source data into the ocean forecast model.

**Alan E. Gelfand**
Duke University
Institute of Statistics and Decision Sciences
*alan@stat.duke.edu*

Principles of Multi-Level Stochastic Modeling

Multi-level stochastic modeling has emerged as an important strategy for describing complex processes. Such explanatory models are developed with regard to both observable and unobservable aspects of the process. Uncertainty is incorporated into

each modeling stage and is propagated through the entire model to effectively capture overall uncertainty with regard to inference.

The presentation will review basic ideas of hierarchical modeling. Examples of hierarchical/multi-level models including spatial and dynamic specifications will be offered. The use of graphical models to develop and interpret assumptions implicit in such modeling will be briefly considered. Inference under such modeling will also be discussed, as well as model fitting along with computation required to implement desired inference.

**Thomas Hou**
California Institute of Technology
Department of Applied Mathematics
*hou@acm.caltech.edu*

Multiscale Analysis and Computation of Flow and Transport in Heterogeneous Media

Many problems of fundamental and practical importance contain multiple scale solutions. Composite materials, flow and transport in porous media, and turbulent flow are examples of this type. Direct numerical simulations of these multiscale problems are extremely difficult due to the range of length scales in the underlying physical problems. In this talk, I will give an overview of the multiscale finite element method and describe some of its applications, including composite materails, wave propagation in random media, convection enhanced diffusion, and flow and transport in heterogeneous porous media. It is important to point out that the multiscale finite element method is designed for problems with many or continuous spectrum of scales without scale separation. Further, we introduce a new multiscale analysis for convection dominated 3-D incompressible flow with multiscale solutions. The main idea is to construct semi-analytic multiscale solutions locally in space and time, and use them to construct the coarse grid approximation to the global multiscale solution. Our multiscale analysis provides an important guideline in designing a multiscale method for computing incompressible flow with multiscale solutions.

**Chuanshu Ji**
University of North Carolina-Chapel Hill
Department of Statistics
*cji@email.unc.edu*

Statistical Approaches for Modelling Microstructures and Computation of Material Properties

The connection between microstructures and material properties is a central issue in computational materials science. Important contributions in this direction, such as finite element methods (FEM) and homogenization theory, are primarily made by applied

mathematicians and physical scientists. In contrast, little has been done to address related statistical aspects, due to the enormous complexity involved in microstructure data and certain implicit links from microstructures to properties. Our work attempts at proposing some statistical models for microstructures, fitting the models by real microstructure data, generating synthetic microstructures from the fitted models and constructing confidence intervals for effective properties calculated by FEM, etc.

**Arthur J. Krener**
University of California
Department of Mathematics
*ajkrener@ucdavis.edu*

Control of Nonlinear Systems

We shall review the simplest methods of controller design for linear systems and their generalizations to nonlinear systems. Topics that will be discussed are linear quadratic Gaussian regulation, linear H-Infinity control and estimation, feedback linearization, backstepping and nonlinear H-Infinity control and estimation.

**Christopher S. Lynch**
Georgia Institute of Technology
Department of Mechanical Engineering
*lynch.admin@me.gatech.edu*

Multiscale Modeling Challenges - Ferroelectric Materials

Ferroelectric materials are used extensively in transducer and actuator applications. These applications range from large Navy sonar systems to medical ultrasonic imaging to active vibration suppression. These materials are readily available in ceramic form (PZT and Barium Titanate) and more recently as single crystals (PMN-xPT and PZN-xPT). The ceramic materials are generally acceptor or donor doped to control the electro-mechanical coupling properties by decreasing of increasing domain wall mobility. This results in non-linear and hysteretic constitutive behavior and a coupling between mechanical and electrical fields in the fracture process. Good phenomenological non-linear and hysteretic constitutive models are needed at the macroscale for efficient implementation in finite element programs. These models have been developed based on concepts from metal plasticity such as yield surfaces and hardening [Landis, Fleck, Huber, McMeeking, etc.]. The coupling behavior arises through phenomena that occur at smaller length scales. This has resulted in the application of techniques from micromechanics in which each grain is modeled as a single crystal with Preisach type switching behavior [Hagood, Hwang, Chen, Lynch, McMeeking, Landis etc.]. The macroscopic behavior is found from the microscopic behavior using various volume averaging techniques. These physically based models are multiaxial and well suited to computationally determining yield surfaces and hardening coefficients for the

macroscopic models. Recently, models have been developed that view the grains of the ceramic as being comprised of volume fractions of crystal variants with evolution laws governing the volume fractions [Landis]. This provides a better model of the underlying material behavior, but the evolution laws still need to be developed. Work on single crystals [Lynch, McLaughlin] has shown that not only do the volume fractions of crystal variants evolve through domain wall motion, there are also field induced phase transformations. Phase field modeling [L.Q. Chen, Zhang etc.] is proving to be an excellent tool for studying domain structure and domain wall motion. This presentation will give an overview of the modeling approaches at the different length scales and discuss some of the challenges faced in modeling this complex class of materials.

**Hideo Mabuchi**
California Institute of Technology
Department of Physics and Control & Dynamical Systems
*hmabuchi@caltech.edu*

Real-time Quantum Feedback Control

It is difficult to imagine sophisticated technology without feedback control, and there is every reason to believe this will continue to be true as we begin to explore design and engineering at the atomic scale. As the systems we seek to control become smaller and smaller, there is an inevitable transition from classical physical behavior to quantum physics; the dynamical equations change, and the novel feature of unavoidable measurement backaction comes into play. In this talk, I will describe our group's ongoing experiments on quantum feedback control of spin degrees of freedom in an ensemble of cold atoms. I will argue that real-time feedback is a crucial tool that allows measurement backaction to be harnessed as a novel form of actuation. I will also discuss our perspective on quantum filtering for state estimation, and argue that quantum feedback provides compelling new motivation for further development of the subject of stochastic nonlinear control.

**Murti V. Salapaka**
Iowa State University
Department of Electrical Engineering
*murti@iastate.edu*

Multi-Objective Robust Control

In most control related engineering objectives the performance of a system is evaluated based on multiple measures that include step response based criteria, frequency domain measures, response due to white noise inputs and other criteria such as zero steady state error in tracking ramp signals. It is also of significant engineering importance to incorporate the uncertainty into the design process that accounts for unmodeled, and under-modeled dynamics. In this talk an input-output methodology will be presented that

provides for the design of controllers that address multi-objective and robustness concerns. The use of optimization techniques in obtaining effective solutions and the challenges faced will be delineated.

**Ralph Smith**
North Carolina State University
Department of Mathematics
*rsmith@ncsu.edu*

Energy Techniques for Multiscale Modeling

In this tutorial, we discuss the development of energy-based models to quantify the hysteresis and constitutive nonlinearities inherent to a broad range of high performance materials. This development exploits a multiscale analysis of the materials and relies on certain stochastic homogenization techniques to achieve the efficiency required for system design and model-based control implementation. In the first step of the development, internal, kinetic and thermal energy relations are constructed at the microscopic or mesoscopic levels to quantify fundamental material properties. We then consider certain physical parameters within the energy relations to be manifestations of underlying distributions rather than constants to obtain macroscopic constitutive relations having a small number of effective parameters. We subsequently employ energy analysis to construct system models based on the nonlinear constitutive relations for a number of deviceswhich exploit advanced material architectures. Finally, we will discuss the development of full and reduced-order approximation techniques which facilitate design and control implementation. Examples will be drawn from problems arising in structural acoustics, high speed milling, deformable mirror design, artificial muscle development, tendon design to minimize earthquake damage, and atomic force microscopy.

**Marina Vannucci**
Texas A&M University
Department of Statistics
*mvannucci@stat.tamu.edu*

Bayesian Inference for Wavelet-Based Modeling of Functional Data

In this talk I will describe Bayesian methodologies for wavelet-based modeling of functional data extended beyond the single curve case. I will first consider the choice of explanatory variables in multivariate linear regression models with predictors arising as curves. I will use wavelet transforms to represent the curves through wavelet coefficient sets describing local features in a parsimonious way. I will then use mixing priors and MCMC methods to select wavelet coefficients that best predict a multivariate response. I will show applications to infrared spectroscopy and also discuss extensions of the methodologies to classification models. Next, I will address the problem of mean function estimation for functional data that have a nested hierarchical structure. I will use

experimental data arising from carcinogen-induced colon cancer in rodent models. Here multiple shrinkage priors on the wavelet coefficients at the top of the hierarchy will result in adaptive regularization of the function estimates. The method will lead to estimates and posterior credible intervals for the mean function and random effects functions, as well as the variance components of the model.

**Yazhen Wang**
University of Connecticut
Department of Statistics
*yzwang@merlot.stat.uconn.edu*

"Interval Estimation of Self-Similarity Index for Locally Self-Similar Processes"

Many naturally occurring phenomena produce data that exhibits self-similar behavior which evolves as the phenomena progress. Adequate modeling of such data requires the consideration of locally self-similar processes with time-varying self-similarity index function. Wavelet based methodology has been developed to estimate the self-similarity index function. This talk will present recent work on constructing confidence intervals for the self-similarity index. We derive an asymptotic distribution for the wavelet-based estimators of self-similarity index and use the distribution results to construct confidence intervals. Numerical simulations and applications are carried. This is joint work with Joe Cavanaugh at University of Iowa and Wade Davis at Baylor University.

**C. F. Jeff Wu**
Georgia Institute of Technology
Department of Industrial and Systems Engineering
*jeffwu@isye.gatech.edu*

"Exploitation and Integration of Detailed and Quick FEA Simulations: Improving Engineering Design via a Bayesian Synthesis"

This talk is motivated by collaborative work on robust topology design of cellular material at Georgia Tech. In simulating the material properties finite elements analysis (FEA) can be done based on different physical-mechanistic models. Typically a more detailed or accurate model will require longer FEA runs while a simplified or rough model will require quicker FEA runs. They are referred to as *detailed* and *quick* simulations respectively. Detailed simulations can take up days of CPU time. While they can provide more accurate results, their number can be limited. On the other hand, many quick simulations can be obtained, though the results are less reliable. A new approach is taken here to combine these sources of data to come up with a *meta-model* that can be used to describe the relationship between the output of FEA runs (i.e., material properties) and input parameters (i.e., design parameters) and for prediction. Since the quick simulations form the bulk of the data, they are used to build a semi-parametric model based on Gaussian random functions. This fitted model is then "adjusted" by

incorporating the information in the detailed simulations. The model adjustment is done using a Bayesian synthesis. Real (but preliminary) data will be used to illustrate this technique. This approach is also applicable to quick and detailed simulations that are based on *fine and coarse scales* of the same physical model. In this sense the reported work can be considered as an example of multiscale modeling and analysis.

(Based on joint work with Z. Qian, R. Joseph, C. Seepersad, J. Allen.)

*C. Opening Workshop Poster List*

- **Dennis D. Cox**, Rice University
  "Stochastic Relaxation of Variational Integrals with Nonattainable Infima"

- **Marcelo J. Dapino**, The Ohio State University
  "New Magnetoelastic Transduction Principle in Ferromagnetic Shape Memory Materials"

- **Boumediene Hamzi**, University of California-Davis
  "The Controlled Center Dynamics"

- **Bongsoo Jang**, University of North Carolina-Charlotte
  "The Calculation of Effective Coefficients"

- **Denise Krueger**, Virginia Polytechnic Institute
  "Stabilized Finite Element Methods and Feedback Control"

- **Don Leo**, Virginia Polytechnic Institute and State University
  "Multiscale Modeling of Ionically-Conductive Membranes as Electromechanical Transducers"

- **Tieqi Liu**, Georgia Institute of Technology
  "Characterization and Modeling of Domain Engineered Relaxor Single Crystals"

- **Hanxiong Luo**, University of California-San Diego
  "Design, Modeling, and Optimization of Compliant Tensegrity Fabrics for the Reduction of Turbulent Skin Friction"

- **Yevgen Melikhov**, Iowa State University
  "Study of Extended Preisach Model for Description of Inverted Hysteresis Loops with Negative Coercivity"

- **William S. Oates**, Georgia Institute of Technology
  "Characterization and Modeling of Domain Engineered Relaxor Single Crystals"

- **Daniel R. Reynolds**, Lawrence Livermore National Laboratory
  Computing Phase Transitions in Shape Memory Alloy Wires

- **John Singler**, Virginia Polytechnic Institute
  Sensitivity of Fluid Flow to Small Perturbations of the Boundary

- **Ramon van Handel**, California Institute of Technology
  Feedback Control of Quantum State Reduction

*D. Workshop on Multiscale Challenges in Soft Matter Materials Program*
February 15-17, 2004

**Sunday – February 15, 2004**
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

**11:45-12:45 pm** Registration and Brunch

**12:45-1:00 pm** Introduction and Welcome
**Jim Berger**, SAMSI and **Greg Forest**, University of North Carolina-Chapel Hill

**1:00-2:00 pm** **Overview**: Experiments and Phenomena
Chair: **Ralph Smith**, North Carolina State University

- **Maria Kilfoil**, McGill University
- **Patrick Mather**, University of Connecticut

**2:00-3:00 pm** **Overview**: Computational Methods and Scales
Chair: **Ralph Smith**, North Carolina State University

- **Mike Graham**, University of Wisconsin-Madison
- **Yannis Kevrekidis**, Princeton University

**3:00-3:30 pm** Coffee Break

**3:30-5:00 pm** **Overview**: Theory and Modeling
Chair: **Chuanshu Ji,** University of North Carolina-Chapel Hill

- **Alex Rey**, McGill University
- **Michael Rubinstein**, University of North Carolina-Chapel Hill
- **Qi Wang**, Florida State University

**6:30-8:30 pm** **Reception and Poster Session** at NISS-SAMSI

(19 T.W. Alexander Drive, 919-685-9350)

***Poster Presenters**:* The Radisson Shuttle will depart at 5:45pm to bring to SAMSI to set-up your poster.

***All other Participants**:* Carolina Livery will be providing continuous shuttle service between the Radisson and SAMSI. The first shuttle will depart from the Radisson at 6:20pm and the last shuttle will leave SAMSI at 8:35pm.

**Monday – February 16, 2004**
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

**8:00-8:30 am** Continental Breakfast

**8:30-9:30 am** **Session 1:** Experiments and Phenomena Part I
Chair: **Maria Kilfoil**, McGill University

- *8:30-8:45* Eugenia Kumacheva
- *8:45-9:00* Anette Hosoi
- *9:00-9:15* Karen Daniels
- *9:15-9:30* Patrick Mather

**9:30-9:45 am** Coffee Break

**9:45-11:00 am** **Session 1:** Experiments and Phenomena Part II
Chair: **Patrick Mather**, University of Connecticut

- *9:45-10:00* Maria Kilfoil
- *10:00-10:15* Linda Smolka
- *10:15-10:30* Stephen Bechtel
- *10:30-10:45* Sergei Sheiko

- *10:45-11:00* Cross-fire Discussion

**11:00-11:15 am** Coffee Break

**11:15-12:15 pm** **Session 2:** Computation Part I
Chair: **Mike Graham**, University of Wisconsin-Madison

- *11:15-11:30* Yannis Kevrekidis
- *11:30-11:45* Peter Mucha
- *11:45-12:00* McKay Hyde
- *12:00-12:15* Weinan E

**12:15-1:15 pm** Lunch, Room A-B (2nd Floor)

**1:15-2:30 pm** **Session 2:** Computation □Part II
Chair: **Yannis Kevrekidis,** Princeton University

- *1:15-1:30* Ruhai Zhou
- *1:30-1:45* Simon Tavener
- *1:45-2:00* Shi Jin
- *2:00-2:15* Bo Li

- *2:15-2:30* Cross-fire Discussion

**2:30-2:45 pm** Coffee Break

**2:45-4:00 pm** **Session 3:** Theory and Modeling □Part I
Chairs: **Alex Rey**, McGill University
**Michael Rubinstein**, University of North Carolina-Chapel Hill
**Qi Wang**, Florida State University

- *2:45-3:00* Alain Goriely
- *3:00-3:15* Roman Grigoriev
- *3:15-3:30* Andrey Dobrynin
- *3:30-3:45* Eliot Fried
- *3:45-4:00* Anne Robertson

**4:00-4:15 pm** Coffee Break

**4:15-5:30 pm** **Session 3:** Theory and Modeling □Part II
Chairs: **Alex Rey**, McGill University
**Michael Rubinstein**, University of North Carolina-Chapel Hill
**Qi Wang**, Florida State University

- *4:15-4:30* Pam Cook
- *4:30-4:45* Peter Constantin
- *4:45-5:00* Leonid Berlyand
- *5:00-5:15* Guillermo Goldsztein

- *5:15-5:30* Cross-fire Discussion

**6:30 pm** **Dinner in Chapel Hill**
Carolina Livery will be providing transportation to University Square on Franklin Street in Chapel Hill. Shuttle service will be 6:30-10:00pm. The shuttle will be leaving the Radisson at 6:30, 7:30, 8:30,

9:30. It will leave University Square at 7:00, 8:00, 9:00, 10:00 to bring participants back to the Radisson.

**Tuesday – February 17, 2004**
Radisson Hotel Research Triangle Park
Room H (3rd Floor)

| | |
|---|---|
| **8:00-9:00 am** | Continental Breakfast |
| **9:00-9:30 am** | **Open Forum**: Identifying Breakout Groups<br>Chair: **Greg Forest,** University of North Carolina-Chapel Hill<br>• The goal is to identify collaborations, research projects, and future challenges requiring theory & modeling, computation and experiments. |
| **9:30-10:30 am** | **Breakout Group Discussions** with summary reports |
| **10:30-11:00 am** | Coffee Break |
| **11:00-11:45 am** | **Breakout Group Reports** |
| **11:45-12:30 pm** | Final Remarks and Discussion of Proceedings<br>• The goal is to target proceedings volume that summarizes the status of experiments, computation, and theory & modeling and describes the next set of challenges and opportunities. |
| **12:30 pm** | Adjournment |

*E. Workshop on Multiscale Challenges in Soft Matter Materials Abstracts*

**Stephen Bechtel**
Ohio State University
Department of Mechanical Engineering
*bechtel.3@osu.edu*

Experiments and Phenomena in Polymer/Nanoparticle Composite Systems

Abstract:
(None submitted)

**Leonid Berlyand**
Penn State University
Department of Mathematics
*berlyand@math.psu.edu*

Ginzburg-Landau minimizers with prescribed degrees in perforated domains. Capacity of the domain and emergence of vortices.

Let Ω be a 2D domain with holes $\omega_0, \omega_1, \ldots, \omega_j, j = 1 \ldots k$. In domain $A = \Omega \setminus \cup_{j=0}^{k} \omega_j$ consider class $\mathcal{J}$ of complex valued maps having degrees 1 and -1 on $\partial\Omega, \partial\omega_0$ respectively and degree 0 on $\partial\omega_j, j = 1 \ldots k$.

We show that if cap($A$) $\geq \pi$, minimizers of the Ginzburg-Landau energy $E_K$ exist for $K$. They are vortexless and converge in $H^1(A)$ to a minimizing $S^1$-valued harmonic map as the coherency length $K^{-1}$ tends to 0. When cap($A$) < $\pi$, we establish existence of quasi-minimizers, which exhibit a different qualitative behavior: they have exactly two zeroes (vortices) rapidly converging to $\partial A$.

**Peter Constantin**
The University of Chicago
Department of Mathematics
const@math.uchicago.edu

Asymptotic States of Smoluchowski Equations

We discuss the strong potential limit of long time behavior of Smoluchowski equations. Issues of regularity, dissipativity, number of states and asymptotic behavior will be mentioned.

**Pam Cook**
University of Delaware
Department of Mathematical Sciences
*cook@math.udel.edu*

Title and Abstract:
(None submitted)

**Karen Daniels**
Duke University
Department of Physics
*ked@phy.duke.edu*

Shearing and Disorder in a Vibrationally Fluidized 3D Granular Flow

Monodisperse granular systems crystallize under vibration and become disordered when sheared. Experiments in a 3D Couette cell vibrated from below and sheared from above show a disordering transition which is coincident with the transition from stick-slip to continuous flow in a shear band. We have characterized the statistical properties of this flow though force distributions, dilation measurements, velocity profiles, cluster statistics, and velocity profiles. As a function of these two means of driving the system, we find that transitions in these diverse measures occur where the shear and vibrational energies are approximately equal.

**Andrey Dobrynin**
University of Connecticut
Institute of Materials Science and Chemical Engineering
*avd@ims.uconn.edu*

Molecular Dynamics Simulations of Layer-by-Layer Polyelectrolyte Self-Assembly

Abstract:
(None submitted)

**Weinan E**
Princeton University
Department of Mathematics
*weinan@math.princeton.edu*

Title and Abstract:
(None submitted)

**Eliot Fried**
Washington University-St. Louis
Department of Mechanical and Aerospace Engineering
*efried@me.wustl.edu*

Universal States In Nematic Elastomers

A state which can be maintained in equilibrium for a particular material model by the action of surface tractions alone is called controllable. If a state is controllable for all material models in a particular class, that state is called universal. Universal states are of central importance in the design of experiments for the determination of constitutive relations, as such experiments should not be based on specific model expressions but, rather, should be produce states sustainable by the full range of material models belong to a broadly relevant class. Universal states have been used with benefit for the study of

elastomeric solids, viscoelastic fluids, and uniaxial nematic liquid crystals. In addition to background material, this talk will discuss recent results and open problems related to universal states in nematic elastomers.

**Guillermo Goldsztein**
Georgia Institute of Technology
Department of Mathematics
*ggold@math.gatech.edu*

Transport of Nutrients in Bones

We study a model bone that captures the main physical effects responsible for the transport of nutrients in real bones under cyclic load. In particular, we obtain the dependence of the rate of transport of nutrients on the microgeometry of the model bone and the applied load.

**Alain Goriely**
University of Arizona
Department of Mathematics
*goriely@math.arizona.edu*

Old and New Challenges in the Modeling of Growth in Biological Systems

In this talk, I will review different ways to model growth of elastic material as applicable to biological systems. Then, I will discuss various challenges and issues related to growth at the experimental, theoretical and computational levels. Since time does not permit, I will only briefly talk about my own (minor) contribution that will be otherwise illustrated in the poster session.

**Mike Graham**
University of Wisconsin-Madison
Department of Chemical and Biological Engineering
*graham@engr.wisc.edu*

Computational Challenges In Flowing Soft Materials

Abstract:
(None submitted)

**Roman Grigoriev**
Georgia Institute of Technology
Department of Physics

*roman.grigoriev@physics.gatech.edu*

Thin Liquid Films: Dynamics, Stability and Control

Abstract:
(None submitted)

**Anette Hosoi**
Massachusetts Institute of Technology
Department of Mechanical Engineering
*peko@mit.edu*

Two-Dimensional Self-Assembled Patterns in Diblock Copolymers

The ability to control patterned structures on the molecular scale by noncovalent forces will be a powerful tool for the miniaturization of devices. We have developed a system in which behavior of diblock copolymers in two dimensions can be optimized to produce regular, uniform features of molecular dimensions. We present a mathematical model for nanoscale pattern formation in diblock copolymers which captures the dynamic evolution of a solution of poly(styrene)-poly(ethyleneoxide) in solvent, PS-PEO, at an air-water interface. The model has no fitting parameters and incorporates the effects of surface tension gradients, evaporation of solvent, entanglement and diffusion. The mechanism for pattern formation differs fundamentally from spinodal decomposition as Van der Waals forces are negligible (relative to polymer entanglement) in this system. The resultant morphologies compare well qualitatively and quantitatively with experimental data.

**E. McKay Hyde**
University of Minnesota
School of Mathematics
*hyde@math.umn.edu*

Fast, High-Order Methods in Computational Electromagnetics and Acoustics

Abstract:
(None submitted)

**Shi Jin**
University of Wisconsin-Madison
Department of Mathematics
*jin@math.wisc.edu*

Numerical Methods for Multiscale Kinetic Problems

I will briefly review several recent methods for kinetic problems where the mean free path has different orders of magnitude. In particular, I will present

1) asymptotic-preserving methods: which solve the kinetic problems with numerical resolution at hydrodynamic scales without using the hydrodynamic equations.

2) domain decomposition methods: we provide interface conditions that allow us to couple a kinetic equation with a (hydrodynamic) diffusion equation for numerical computation without using iterations at each time step.

**Yannis Kevrekidis**
Princeton University
Department of Chemical Engineering
*yannis@princeton.edu*

Title and Abstract:
(None submitted)

**Maria Kilfoil**
McGill University
Department of Physics
*kilfoil@physics.mcgill.ca*

Overview Talk on Sunday:
Unifying Aspects of Soft Matter Research

Here I will discuss the difficulty in categorizing soft materials in simple terms relating to any sort of common structural or dynamical properties, simply because they are so rich and varied. Instead, they may be categorized by the similar phenomena they exhibit across a broad range of systems. One of the unifying aspects of this research is the search for new ways of solving increasingly complex equations or models that can be written down for different systems.

Short Talk:
In Search of Quantification of Disorder in Amorphous Colloidal Gels

Abstract:
(None submitted)

**Eugenia Kumachev**
University of Toronto
Department of Chemistry
*ekumache@chem.utoronto.ca*

Colloid Crystallization in Constrained Geometry

We describe a non-equilibrium, convective, mechanism leading to formation of ordered 2D structures of both closed-packed hexagonal and non-closed-packed rhombic symmetries. The number and types of possible lattices is determined by the ratio of the width of the channel to the diameter of the particle. The structures tend to return to a regular lattice after a defect is introduced; that is, for example, they tend to self-repair disorder induced by particle polydispersity, contaminants, and flow instabilities. The stability of different lattices is analyzed numerically for particles with different polydispersity.

**Bo Li**
University of Maryland
Department of Mathematics
*bli@math.umd.edu*

Numerical Simulation of Pitaxial Growth of Thin Films

In this talk, we present an adaptive finite element method for the simulation of epitaxial growth of thin films with the attachment-detachment kinetics and the edge-adatom diffusion.

**Patrick Mather**
University of Connecticut
Institute of Materials Science and Chemical Engineering
*mather@ims.uconn.edu*

Overview Talk on Sunday:
Trends in Experimental Studies of Soft Materials

I will overview new experimental studies of soft materials and complex fluids with an emphasis on opportunities for theoretical and computational study. Particular attention will be given to ordered soft materials and their interaction with external forces. Liquid crystals, polymeric gels, and natural materials will be highlighted.

Short Talk:
Smectic Liquid Crystalline Elastomers

It was long expected and recently shown that main-chain liquid crystalline elastomers (MC-LCEs) may serve as high performance soft actuators due to a coupling of their intrinsic characteristics of high, yet labile, ordering and network strain. Here, we present new siloxane-based smectic MC-LCEs. These new materials exhibit a unique thermomechanical behavior known as shape memory effect; that is, the ability to adopt a prescribed temporary shape by fixing and subsequently recover to the equilibrium shape

under external stimulus. We will hypothesize on the underlying molecular phenomenon responsible for the shape memory behavior exhibited by the new siloxane-based smectic liquid crystalline elastomers.

Joint work with Ingrid A. Rousseau

**Peter J. Mucha**
Georgia Institute of Technology
Department of Mathematics
*mucha@math.gatech.edu*

Physical Applications of Interacting Particle Systems

A number of interesting physical systems can be well approximated by a description in terms of direct interactions between particles. Indeed, while the microscopic interparticle rules are well known in many cases, the description in terms of those rules is often more complicated than is desired or even possible to compute. A reduced order or continuum description is then preferable, if such a model can be found. Combining considerations from theory, experimental evidence, and computational efficiency, model interacting particle systems can be proposed and simulated, so that reduced descriptions can be tested. In this brief talk, we will highlight recent work on suspensions and granular media.

**Alejandro Rey**
McGill University
Department of Chemical Engineering
*alejandro.rey@mcgill.ca*

Title and Abstract:
(None submitted)

**Michael Rubinstein**
University of North Carolina
Department of Chemistry
mr@unc.edu

Title and Abstract:
(None submitted)

**Sergio Sheiko**
University of North Carolina
Department of Chemistry
*sergei@email.unc.edu*

Visualization: Spreading Kinematics and Dynamics

Abstract:
(None submitted)

**Linda Smolka**
Duke University
Department of Mathematics
*smolka*@math.duke.edu

An Exact Solution for the Extensional Flow of a Viscoelastic Fluid

We consider the free boundary problem of an axisymmteric, cylindrical liquid filament stretching in an extensional flow through a quiescent fluid of negligible viscosity. Our approach provides a systematic framework in which the known Newtonian solution can be generalized to various viscoelastic constitutive models using a condition for the existence of cylindrical solutions. By assuming a power series expansion for the stress, we obtain an analytic solution that describes the filament motion for a viscoelastic filament. We examine this solution in the weakly and strongly viscoelastic limits, as well as in the transient and long time limits. Comparisons of this exact solution with experimental measurements using a viscoelastic polymer solution show strong quantitative agreement. As $t \rightarrow \infty$, both the solution and the observations scale in the Newtonian limit. This transition from viscoelastic to Newtonian scaling provides insight as to how the molecular dynamics of the polymer couple to the filament's motion. This is joint work with Thomas Witelski (Duke), and Andrew Belmonte and Diane Henderson (Penn State).

**Simon John Tavener**
Colorado State University
Department of Mathematics
*tavener@math.colostate.edu*

Numerical Bifurcation Approach to Nonlinear Phenomena

Nonlinear phenomena arising in a range materials science applications can be studied efficiently using a numerical bifurcation approach. I will outline the strengths and weaknesses of this approach using a number of case studies. I will then describe future directions.

**Qi Wang**
Florida State University
Department of Mathematics

*wang@math.fsu.edu*

Overview Talk on Sunday:
"Constitutive Modeling of Complex Fluids of Flexible and Nematic Polymers"

Abstract:
(None submitted)

**Ruhai Zhou**
University of North Carolina
Department of Mathematics
*ruhai@amath.unc.edu*

"High Order Numerical Computations Of Sheared Nematic Polymers"

The Doi kinetic theory with a Marrucci-Greco distortional elasticity potential is the leading model for spatio-temporal structures created in plane shear flow of nematic liquid crystalline polymers (LCPs). The spatio-temporal structure of the polymer varies significantly and the integration of the Smoluchowski equation is computationally challenging. In our numerical method, we first handle the variables for orientation by expanding the orientational distribution function in spherical harmonics. The Smoluchowski equation is thus reduced to a set of partial differential equations in time and space. Then we discretize the spatial variables using high-order finite differences. Adaptive grid generation techniques are implemented. To provide an accurate and stable integration, we employ the spectral deferred correction algorithm. Some examples of numerical simulations are finally present.

Joint work with M. Gregory Forest (UNC-CH), and Qi Wang (FSU)

### *F. Workshop on Multiscale Challenges in Soft Matter Materials Posters*

- **Leonid Berlyand**, Penn State University
  "Discrete Network Approximation for Highly Packed Random Suspensions"

- **Eric Choate**, University of North Carolina
  "Monodomain Response of Arbitrary Aspect Ratio Nematic Polymers in General Linear Planar Flows"

- **Christopher L. Cox**, Clemson University
  "The CAEFF Integrated Model Software Package for Polymer Process Simulation"

- **Andrey Dobrynin,** University of Connecticut
  "Molecular Simulations of Multilayer Electrostatic Self-Assembly"

- **Eliot Fried**, Washington University-St. Louis
  Deformation-induced Biaxial Disclinated States of Strength +1 in a Nematic Elastomer

- **Eugene Gartland**, Kent State University
  Numerical Modeling of Periodic Structures in Liquid Crystal Films

- **Guillermo Goldsztein**, Georgia Institute of Technology
  Effective Properties of Heterogeneous Materials

- **Zhi-Feng Huang**, Florida State University
  Grain Boundary Motion of Lamellar

- **McKay Hyde**, University of Minnesota
  Fast, High-Order Methods in Computational Electromagnetics and Acoustics

- **Maria Kilfoil**, McGill University
  Local Measurement of Compressibility of Low Density Gels Formed in Attractive Systems

- **Joo Hee Lee**, University of North Carolina
  2D Stochastic Simulations of Shear Flow of Liquid-Crystalline Polymers

- **Tiejun Li**, Peking University
  Some Theoretical Results for the Equations of Complex Fluids

- **Louis Madsen**, University of North Carolina
  NMR Studies on Ideal Rodlike and Bent-Core Nematic Liquid Crystals

- **Patrick Mather**, University of Connecticut
  "The Nematic-Isotropic Interface"

- **Peter J. Mucha**, Georgia Institute of Technology
  Diffusivities and Front Propagation in Sedimentation

- **T. Samulski**, University of North Carolina
  NASA University Research Engineering & Technology Institute on Biologically-Inspired Materials

- **Amy Shen**, Washington University-St. Louis
  Dynamics of Granular Chains

- **Simon John Tavener**, Colorado State University
  Numerical Bifurcation with Multiple Physics and Multiple Scales

- **Jesenko Vukadinovic**, University of Wisconsin-Madison
  Regularity and Dissipativity in Gevrey Classes for Solutions of a Smoluchowski Equation

- **Noel J. Walkington**, Carnegie Mellon University
  Numerical Approximation of the Ericksen Leslie Equations

- **Qi Wang**, Florida State University
  Structure Formation and Evolution in Flows of Nematic Liquid Crystal Polymers

- **Peng Yu**, Penn State University
  A New Moment-Closure Approximation to the FENE Model of Polymeric Fluids

- **Xiaoyu Zheng**, University of North Carolina
  Nano-Composite Material Properties: Homogenization Over Flow-Induced Orientational Distribution

- **Hong Zhou**, University of California-Santa Cruz
  Structure Scaling Properties of Confined Nematic Polymers in Plane Couette Cells

*G. Workshop on Fluctuations and Continuum Equations for Granular Flow Program*

**Friday – April 16, 2004**
NISS-SAMSI Building, Room 104

| | |
|---|---|
| **8:20 am** | ***Carolina Livery Shuttle departs from the Radisson for SAMSI*** |
| **8:30-8:55 am** | Registration and Continental Breakfast |
| **8:55-9:00 am** | Opening ceremonies |
| **9:00-10:00 am** | First Session on Statistical Models<br>Chair: **Robert Behringer**, Duke University |
| 9:00-9:30 | *Granular Fluid Dynamics*<br>**Mark Shattuck**, City College of New York |
| 9:30-10:00 | *Network Approximation for Effective Properties of Highly Concentrated Random Suspensions*<br>**Leonid Berlyand**, Pennsylvania State University |

**10:00-10:45 am** Coffee Break

**10:45-12:15 pm** Session on Fluctuations and Transitions
Chair: **Eric Clément**, University of Paris 6

10:45-11:15 □*Generalized Granular Temperature*□
**Lou Kondic**, New Jersey Institute of Technology

11:15-11:45 □*Granular temperature in Shear Granular Flows*□
**Anael Lemaitre**, University of California-Santa Barbara

11:45-12:15 □*Order Parameter Description of Dense Granular Flows*□
**Igor Aronson**, Argonne National Laboratory

**12:15-1:30 pm** Lunch

**1:30-3:00 pm** Discussion Sessions *****

1:30-1:40 10-minute description of semesters of concentration at SAMSI
**H.T. Banks**, North Carolina State University & SAMSI

**3:00-3:30 pm** Coffee Break

**3:30-5:00 pm** Session on Related Systems
Chair: **J. Michael Rotter**, University of Edinburgh

3:30-4:00 *Title TBA*
**Karin Dahmen**, University of Illinois at Urbana-Champaign

4:00-4:30 □*Erosion Patterns and Avalanche Flows on a Sediment Layer*□
**Eric Clément**, University of Paris 6

4:30-5:00 □*The Glass Transition in Hard Spheres and Beyond*□
**Matthias Sperl**, Technical University of Munich

**5:00-6:30 pm** Informal reception

**6:35 pm** ***Carolina Livery Shuttle departs NISS-SAMSI for Radisson***

**Saturday – April 17, 2004**
NISS-SAMSI Building, Room 104

| | |
|---|---|
| **8:20 am** | ***Carolina Livery Shuttle departs from the Radisson for SAMSI*** |
| **8:30-9:00 am** | Continental Breakfast |
| **9:00-10:00 am** | Session on Applications<br>Chair: **Susan Coppersmith**, University of Wisconsin |
| 9:00-9:30 | *The Implications of Fluctuations in Granular Forces and Flows in Silos*<br>**J. Michael Rotter**, University of Edinburgh |
| 9:30-10:00 | *Granular Flow in Silos Observations and Comments*<br>**Jørgen Nielsen**, By og Byg (Danish Building and Urban Research) |
| **10:00-10:45 am** | Coffee Break |
| **10:45-12:15 pm** | Second Session on Statistical Models<br>Chair: **David Schaeffer**, Duke University |
| 10:45-11:15 | *Modeling the Mechanical Behavior of Granular Materials*<br>**James Jenkins**, Cornell University |
| 11:15-11:45 | *Shear Banding in Granular Materials*<br>**Stefan Luding**, Delft University of Technology |
| 11:45-12:15 | *Title TBA*<br>**Robert Behringer**, Duke University |
| **12:15-1:30 pm** | Lunch |
| **1:30-2:59 pm** | Discussion sessions * |
| **2:59-3:00 pm** | Closing ceremony |
| **3:10 pm** | ***Carolina Livery Shuttle departs NISS-SAMSI for Radisson*** |

* During these discussion sessions, we expect to divide into two or more subgroups, presumably focusing on different aspects of planning for a semester of concentration at SAMSI. Besides scientific issues, we might discuss whom should be invited, what

related fields might enrich the program and the big one do the potential rewards justify support of a program in granular materials by a statistics institute?

### *H. Workshop on Fluctuations and Continuum Equations for Granular Flow Abstracts*

**Igor Aronson**
Argonne National Laboratory
Materials Science Division
*aronson@msd.anl.gov*

Order Parameter Description of Dense Granular Flows

This talk will be focused on various aspects of continuum description of shear granular flows, such as avalanches, shear bands, and stick-slips. In contrast to dilute rapid granular flows, the description of slow dense flows presents a significant challenge for theorists. We will overview various continuum models of dense granular matter and illustrate our order parameter description of partially fluidized shear flows.

**Robert Behringer**
Duke University
Department of Physics
bob@phy.duke.edu

No title and abstract submitted

**Leonid Berlyand**
Penn State University
Department of Mathematics and Materials Research Institute
*berlyand@math.psu.edu*

Network Approximation for Effective Properties of Highly Concentrated Random Suspensions

We present a new approach for calculation of effective viscosity of highly packed suspensions of rigid particles in a Newtonian fluid and provide its rigorous mathematical justification.
The main idea of this approach is the reduction of the original continuum problem, which is described by PDE with rough coefficients, to a discrete random network.

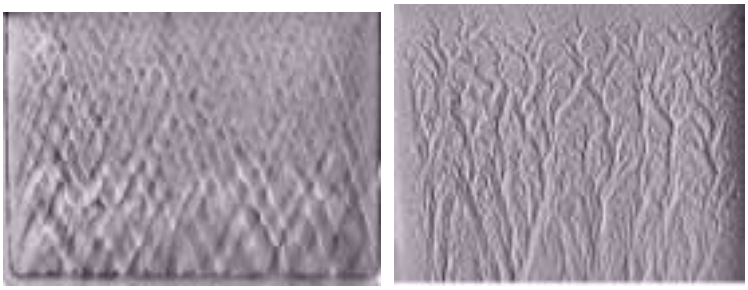
Our mathematical analysis provides an explanation of recent numerical results where the effective viscosity of random suspension of solid particles in a Newtonian fluid was found to be different from earlier predictions based on the lubrication approximation.

We address the issues of geometry of random array of particles and misalignment between the orientation of the particles network and the applied external boundary conditions.

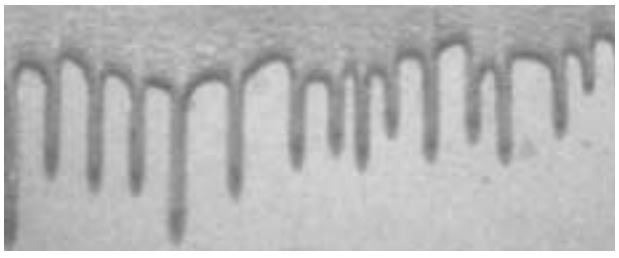
**Eric Clément**
University of Paris 6
erc@ccr.jussieu.fr

□Erosion Patterns and Avalanche Flows on a Sediment Layer□

Many natural patterns coming from erosion have been described by geomorphologists. Erosive patterns are common in situations implying dense sediment transport, but so far, a complete understanding of all the physical processes involved is still a challenging issue. We report here on a laboratory-scale, experiments which reproduce a rich variety of natural patterns with few control parameters. In particular, we produce various rhomboidal structures and ramified networks often found on sandy shores and flats. It turns out that the standard views based on surface waves fall short to explain the phenomenon. We argue that the small thickness of the flowing layer at the onset of grain carriage instead leads to a strongly non-linear erosion-deposition process. From the experimental parameters, the relevance of some possible erosion mechanisms can be analyzed.



Also using the same experimental set-up we are able to produce underwater avalanches instabilities that may display fingering instability of the front very similar in shape to the classical fingering instability found for thin viscous films. We discuss the mechanisms that may lead to the instability propose and analysis of the wave selection dynamics.



Co-workers : Florent malloggi, Adian Daerr, Jose Lanuza

**Karin Dahmen**
University of Illinois at Urbana-Champaign
Physics Department
dahmen@mail.physics.uiuc.edu

No title and abstract submitted

**James T. Jenkins**
Cornell University
Department of Theoretical and Applied Mechanics
*jtj2@cornell.edu*

Modeling the Mechanical Behavior of Granular Materials

We will focus on the quasi-static, rate-independent behavior of granular materials and review recent theories for their deformation and failure. These involve incremental relations between stress and deformation that incorporate fluctuations in the translation and rotations of the particles and include a dependence on the state of the material. The challenge is to identify the appropriate state variables from the consideration of force and moment equilibrium of the individual particles, to incorporate them through suitable averaging into the incremental stress-strain relation, and to describe their evolution with the deformation.

**Lou Kondic** and Robert P. Behringer
New Jersey Institute of Technology
Department of Mathematics
*kondic@m.njit.edu*

Generalized Granular Temperature

We consider the role of elastic energy in the context of granular materials undergoing shear flow. Depending on the ratio of pressure to Young's modulus of the material from which grains are made and the typical velocity of shearing, there is a transition from a regime in which the fluctuations of kinetic energy are dominant to the regime where the fluctuations of elastic energy are dominant. This regime has likely been reached in recent experiments. We then consider a generalization of the granular temperature that includes both types of energy fluctuations and that changes smoothly from one regime to the other. This generalization is then compared to the temperature resulting from a fluctuation-dissipation relation based on equilibrium statistical mechanics. We conclude by discussing an analogy of heat conduction that involves this newly proposed temperature.

**Anael Lemaitre**
University of California, Santa Barbara
Department of Physics
*anael@lob.physics.ucsb.edu*

Granular temperature in Shear Granular Flows

Contact dynamics simulation of sheared granular materials in different geometries are implemented to test recent theoretical ideas and a set of constitutive equations based on the Shear Transformation Zone theory of plasticity in amorphous solids. It was proposed that the collision frequency, determined by a granular temperature, sets the scale of activated rearrangements thereby leading to the observation of Bagnold scaling in dense flows.

A comparative study of (i) sheared granular material in Lees-Edwards geometry with SLLOD equations of motion and (ii) flow down an inclined plane permits us to test the validity of a local rheology assumption, and allows us to measure the parameters of the theory in the former geometry and extrapolate these measurements to the latter. The remarkable agreement shows that the theory can be predictive.

**Stefan Luding**
DelftChemTech
Particle Technology
*s.luding@tnw.tudelft.nl*

Shear Banding In Granular Materials

We present experiments along with molecular dynamics (MD) simulations of 2D and 3D shear cells undergoing slow shearing. The simulation results are compared to experimental studies and good quantitative agreement is found typically.

One aim of the talk is to discuss the micro-macro transition for these systems and to present data on the stress- and deformation-tensors under shear. The eventual goal is to obtain constitutive relations for the behavior of compressible, anisotropic granular media under slow shear. Finally, also the rotation of the particles and their frictional interaction is of interest.

**Jørgen Nielsen**
Danish Building and Urban Research
*jn@dbur.dk*

Granular Flow In Silos - Observations And Comments

The talk will focus on the statement that flow in silos is related to many phenomena, which challenge the principles of Continuums Mechanics and the deterministic treatment of those phenomena.

Examples of such phenomena, where the principles of Continuum Mechanics fail, are the onset of flow through an outlet and the pressure distributions on areas in cases where the

cross-sectional dimensions of the outlets or the areas are of the same order of magnitude as the particle diameter.

Difficulties with the deterministic treatment originate from many causes that are not fully understood, and from lack of knowledge concerning the values of the equivalent continuum material parameters, which may apply to the materials stored in the silo in the future. This makes the mathematical models incomplete and thus inaccurate. Furthermore, the load models that are used in design standards must be simple so that they do not reflect the pressure fluctuations, which are seen in real silos.

The talk will include some comments concerning scaling laws, the limitation of the Continuum Mechanics approach, and the statistical treatments of silo phenomena.

**J. Michael Rotter**
University of Edinburgh
M.Rotter@ed.ac.uk

□The Implications of Fluctuations in Granular Forces and Flows in Silos□

This talk will relate scientific explorations of granular solids packing structures, force distributions and flow variations to their engineering implications in the design of silos for the storage of large volumes of solids. The aim of the talk is to provide a focus on many of the key questions that need answers to advance the technology of granular solids storage. It will discuss conditions in which a statistical treatment is necessary and those where a deterministic treatment is adequate, and it will explore the appropriate roles of discrete particle and continuum models in predicting silo pressures. Both spatial and time-dependent fluctuations will be considered.

As its starting point, the talk begins with the consequences for the silo structure of pressure fluctuations on a silo wall. These fluctuations may be both temporal and spatial, but the latter are evidently more demanding. The manner in which structural safety calculations are performed using approximate statistical evaluations of experimental observations is outlined, showing that there is great scope for scientific work to improve the predictive models. These could be done either using continuum treatments involving stochastic property variations or explorations of deterministic phenomena that lead to spatial variations in pressure patterns on a large scale.

The problem of the uncertainty of pressure observations will be noted, in the context of stochastic variations in discrete particle forces and their relevance to the interpretation of silo pressure data. The avoidance of this problem by direct measurement of stresses developed in the structure is shown to be helpful but does not resolve the difficulties.

The talk is intended to provide food for thought on the very many ways in which statistical treatments using continuum models should be able to advance our understanding and practical exploitation of the science of granular solids in silos.

**Mark Shattuck**
City College of New York
Department of Physics
*shattuck@ccny.cuny.edu*

Granular Fluid Dynamics

Imagine a world where gravity is so strong that if an ice cube is tilted the shear forces melt the surface and water avalanches down. Further imagine that the ambient temperature is so low that the water re-freezes almost immediately. This is the world of granular flows. As a granular solid is tilted the surface undergoes a sublimation phase transition and a granular gas avalanches down the surface, but the inelastic collisions rapidly remove energy from the flow lowering the granular temperature (kinetic energy per particle) until the gas solidifies again. It is under these extreme conditions that we attempt to uncover continuum granular flow properties. Typical continuum theories like Navier-Stokes equation for fluids follow the space-time evolution of the first few moments of the velocity distribution. We study continuously avalanching flow in a rotating two-dimensional granular drum using high-speed video imaging and extract the position and velocities of the particles. We find a universal near Gaussian velocity distribution throughout the flowing regions, which are characterized by a liquid-like radial distribution function. In the remaining regions, in which the radial distribution function develops sharp crystalline peaks, the velocity distribution has a Gaussian peak but is much broader in the tails.
In a companion experiment on a vibrated two-dimensional granular fluid under constant pressure, we find a clear gas-solid phase transition in which both the temperature and density change discontinuously. This suggests that a low temperature crystal and a high temperature gas can coexist in steady state. This coexistence could result in a narrower, cooler, Gaussian peak and a broader, warmer, Gaussian tail like the non-Gaussian behavior seen in the crystalline portions of the rotating drum.

**Matthias Sperl**
Technical University of Munich
Department of Physics
*msperl@ph.tum.de*

The Glass Transition In Hard Spheres And Beyond

I will briefly review the statistical physics approach applied to the glass transition in colloidal systems. The hard-sphere system, where only excluded volume determines the interaction, displays a glass transition that is well established in experiments and computer-simulation studies and is also understood theoretically to a reasonable degree.

In recent years, the results have been extended to binary mixtures of hard spheres, non-spherical particles with hard-core repulsion, and hard spheres with short-ranged attraction. Subtle and to some extent non-trivial packing effects are found when comparing the glass transitions in these more complex systems with the glass transition in the pure hard-sphere system. In many cases, theoretical predictions could be tested against experiments and computer simulation, providing insight into the rich phenomenology of colloidal glasses.

Similar packing effects as in colloids have been found for dense granular systems that resemble hard-sphere mixtures and non-spherical hard-core particles, thus motivating a number of questions to be raised in conclusion.

## VII. Education and Outreach Programs

### *A. Interdisciplinary Workshop for Undergraduates*
June 9-13, 2003

**Monday – June 9, 2003**
NISS-SAMSI Building, Room 104

**8:00 am** Vans Pickup at Sullivan Hall for trip to SAMSI

**9:00-12:00 pm** Presentation by SAMSI Stochastic Computation Group

*9:00-9:35* Contingency Tables
**Mark Huber** & **Ian Dinwoodie**, Duke University

*9:40-10:15* Computation in large-scale graphical models with applications in genomics
**Adrian Dobra** & **Chris Hans**, Duke University

*10: 15-10:45* BREAK

*10:45-11:20* Model Selection
**Rui Paolo**, SAMSI & NISS

*11:25-12:00* Financial Models
**German Molina**, SAMSI & Duke University

**12:00-1:00 pm** Lunch

**1:00-4:00 pm** Presentations by SAMSI Environmental Modeling Group

*1:00-1:30* Statistics of Climate Change
**Richard Smith**, University of North Carolina

| | |
|---|---|
| *1:30-2:00* | Climate Extremes<br>**Amy Grady**, NISS |
| *2:00-2:30* | Turbulent diffusion: Mixing and Extrainment<br>**Richard McLaughlin**, University of North Carolina |
| *2:30-3:00* | BREAK |
| *3:00-3:30* | Modeling Porous Medium Systems<br>**Casey Miller**, University of North Carolina |
| *3:30-4:00* | Optimization and Control of Subsurface flow and transport<br>**C. Tim Kelley**, North Carolina State University |

**4:30 pm** Vans leave SAMSI for cookout at Lake Crabtree

**5:00 pm** Cookout at Lake Crabtree

**Tuesday – June 10, 2003**

| | |
|---|---|
| **9:00-10:30 am** | Introduction to the Forward Problem: The *Vibrating Beam* Application<br>**Jeffrey Hood**, North Carolina State University and SAMSI |
| **10:45-12:00 pm** | Introduction to MATLAB with an IDE Solver Tutorial<br>**Jeffrey Hood**, North Carolina State University and SAMSI |
| **12:00-1:30 pm** | LUNCH |
| **1:30-2:15 pm** | Basic Statistical Concepts<br>**Danny Walsh**, SAMSI |
| **2:30-3:15 pm** | Some Essentials about Probability<br>**Yanyuan Ma**, North Carolina State University and SAMSI |
| **3:30-5:00 pm** | Plot the Data!!! A MATLAB Tutorial<br>**Yanyuan Ma**, North Carolina State University and SAMSI<br>**Danny Walsh**, SAMSI |

**Wednesday – June 11, 2003**

**9:00-11:00 am** *Vibrating Beam* Data Collection at the CRSC Laboratory

| | |
|---|---|
| | **Brandy Benedict**, North Carolina State University |
| **11:00-12:30 pm** | LUNCH |
| **12:30-1:15 pm** | Reflection on the Data Collection Experience<br>**Karen Chiswell**, North Carolina State University and SAMSI |
| **1:30-3:15 pm** | Statistical Estimation in Practice. A MATLAB Tutorial.<br>**Karen Chiswell**, North Carolina State University and SAMSI |
| **3:15-5:00 pm** | Linear Inverse Problems. A MATLAB Tutorial.<br>**Johnathan Bardsley**, North Carolina State University and SAMSI |

**Thursday – June 12, 2003**

| | |
|---|---|
| **9:00-9:45 am** | Nonlinear Inverse Problems: The *Vibrating Beam* Application<br>**Johnathan Bardsley**, North Carolina State University and SAMSI |
| **10:00-10:30 am** | Solving the *Vibrating Beam* Inverse Problem<br>**Brandy Benedict**, North Carolina State University |
| **10:30-11:00 am** | Data Error Analysis: The *Vibrating Beam* Application<br>**Karen Chiswell**, North Carolina State University and SAMSI |
| **11:00-12:00 pm** | Work on the Inverse Problem |
| **12:00-1:30 pm** | LUNCH |
| **1:30-4:30 pm** | Work on the Inverse Problem and Discussion |
| **5:00 pm** | Durham Bulls Baseball |

**Friday – June 13, 2003**

| | |
|---|---|
| **9:30-12:00 pm** | Discussion of Results |
| **12:00 pm** | LUNCH & Adjournment |

*B. CRSC/SAMSI Industrial Mathematical and Statistical Modeling Workshop for Graduates*
July 21-29, 2003

**Sunday – July 20, 2003**

Arrival of Participants at North Carolina State University

**Monday – July 21, 2003**

| | |
|---|---|
| **7:00-8:00 am** | Breakfast in the cafeteria |
| **8:15 am** | Professor Negash Medhin will guide you to Harrelson Hall |
| **8:30-9:00 am** | Coffee and soft drinks in Harrelson 245 |
| **9:00-12:30 pm** | Presentation of Problems |
| **12:30-1:30 pm** | Lunch in cafeteria |
| **1:30-5:00 pm** | Working Session |
| **5:30-7:00 pm** | Pizza Party at Two Guys Restaurant |

**Tuesday – July 22, 2003**

All-day Working Session

**Wednesday – July 23, 2003**

All-day Working Session

**Thursday – July 24, 2003**

All-day Working Session

**Friday – July 25, 2003**

| | |
|---|---|
| **AM** | Working Session |
| **PM** | Centennial Campus & Math Lab Tour |
| **5:30 pm** | Van pickup for Durham Bulls Baseball Game |

**Saturday – July 26, 2003**

**AM** Working Session

**11:30 am** Van pickup at dorm for picnic at Lake Crabtree

**Sunday – July 27, 2003**

**10:30 am** Brunch at the Talley Center

FREE AFTERNOON

**Monday – July 28, 2003**

All-day Working Session

**Tuesday – July 29, 2003**

**AM** Working Session

**1:30 pm** Formal Results Presentation

**6:00-9:00 pm** Dinner at the University Club

**Wednesday – July 30, 2003**

Adjournment and Participant Departure

C. *Undergraduate Workshop on Data Mining: Handling the Flood of Data*
November 14-15, 2003

**Friday – November 14, 2003**
NISS-SAMSI Building, Room 104

**9:30-10:00 am** Continental Breakfast

**10:00-10:15 am** Welcome and Introductions

**10:15-12:30 pm** **Alan Karr**, Director of NISS & Associate Director of SAMSI

*10:15-10:30* About SAMSI
*10:30-10:45* About NISS
*10: 45-12:30* Introduction to Data Mining

**12:30-1:15 pm** Lunch

**1:15-2:45 pm** Mining Software Engineering Data
**Ashish Sanil**, NISS Research Statistician

**2:45-3:00 pm** Break

**3:00-4:30 pm** Mining Pharmaceutical Data
**Stanley Young**, NISS Assistant Director for Bioinformatics

**4:30 pm** Adjourn for the Day

**5:30 pm** Pizza Party at Wellesley Inn

**Saturday – November 15, 2003**
NISS-SAMSI Building, Room 104

**9:30-10:00 am** Continental Breakfast

**10:00-11:00 am** Starting to Mine Real Data
**Alan Karr**

**11:00-12:00 pm** Related Problems that Pose Challenges to DM: Data Confidentiality and Data Quality
**Alan Karr**

**12:00 pm** Adjourn

*D. Undergraduate Workshop on Data Mining: Handling the Flood of Data*
February 13-14, 2004

**Friday – February 13, 2004**
NISS-SAMSI Building, Room 104

**10:00-10:15 am** Welcome and Introductions
**H.T. Banks**, Director of CRSC and Associate Director of SAMSI

**10:15-12:30 pm** **Alan Karr**, Director of NISS & Associate Director of SAMSI

*10:15-10:30* About SAMSI
*10:30-10:45* About NISS
*10: 45-12:30* Introduction to Data Mining

**12:30-1:15 pm** Lunch

| | |
|---|---|
| **1:15-2:45 pm** | Mining Software Engineering Data<br>**Ashish Sanil**, NISS Research Statistician |
| **2:45-3:00 pm** | Break |
| **3:00-4:30 pm** | Mining Pharmaceutical Data<br>**Stanley Young**, NISS Assistant Director for Bioinformatics |
| **4:30 pm** | Adjourn for the Day |
| **5:30 pm** | Pizza Party at Wellesley Inn |

**Saturday – February 14, 2004**
NISS-SAMSI Building, Room 104

| | |
|---|---|
| **10:00-11:00 am** | Starting to Mine Real Data<br>**Alan Karr** |
| **11:00-12:00 pm** | Related Problems that Pose Challenges to DM: Data Confidentiality and Data Quality<br>**Alan Karr** |
| **12:00 pm** | Adjourn |

VIII. Other SAMSI Events

*A. Hot Topics Workshop on Mathematical Sciences Research to Meet National Security Needs*
NISS-SAMSI Building
April 1-2, 2004

**Thursday – April 1, 2004**

| | |
|---|---|
| **8:30-9:00 am** | Continental Breakfast |
| **9:00-9:30 am** | Welcome and Introductions<br>**Jim Berger**, SAMSI<br>**Alan Karr**, NISS and SAMSI |
| **9:30-10:15 am** | CDC Perspective<br>**Lawrence Cox**, National Center for Health Statistics |
| **10:15-10:45 am** | DARPA Perspective<br>**Douglas Cochran**, DARPA |

| | |
|---|---|
| **10:45-11:00 am** | Break |
| **11:00-11:30 am** | DoD Perspective<br>**Nancy Spruill**, Department of Defense |
| **11:30-12:00 pm** | NSA Perspective<br>**William Szewcyzk**, National Security Agency |
| **12:00-12:30 pm** | Agroterrorism Perspective<br>**Barrett Slenning**, North Carolina State University |
| **12:30-1:30 pm** | Lunch |
| **1:30-3:00 pm** | Two-Minute Madness, General Discussion, Formation of Working Groups |
| **3:00-3:30 pm** | Break |
| **3:30-5:30 pm** | Working Groups |

**Friday – April 2, 2004**

| | |
|---|---|
| **8:30-9:00 am** | Continental Breakfast |
| **9:00-10:30 am** | Working Group Summary Preparation |
| **10:30-11:00 am** | Break |
| **11:00-12:30 pm** | Working Group Reports and Discussion |
| **12:30-1:30 pm** | Lunch |
| **1:30-3:00 pm** | Concluding Panel Discussion<br>• Completion of White Paper<br>• Possible SAMSI Program<br>• Other Next Steps |

**APPENDIX E – Workshop Evaluation Summaries**

Workshop participants were given an evaluation questionnaire to complete in each of the SAMSI workshops. A sample questionnaire is given on the following pages.

Below are the summaries of the participant evaluations for the four main scientific workshops held to date. The rating scale was 1-5 (lowest to highest). The five questions addressed in the table were:

a. Scientific Quality
b. Staff Helpfulness
c. Meeting Room/AV Facilities
d. Lodging
e. Local Transportation

**Data Mining Workshop**
**Alan Karr, Leader**
**September 6-10, 2003**
**Total Responses: 37**
**Total Participants: 105**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 1 | 10 | 25 | 1 |
| Staff | 0 | 0 | 0 | 4 | 32 | 1 |
| Facilities | 0 | 0 | 1 | 10 | 25 | 1 |
| Lodging | 0 | 0 | 0 | 13 | 15 | 9 |
| Transport | 0 | 0 | 1 | 12 | 15 | 9 |

**Data Mining Workday**
**Support Vector Machines**
**Marc Genton, Leader**
**January 28, 2004**
**Total Responses:**
**Total Participants:**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | No Formal Registration or Evaluation | | | | | |
| Staff | | | | | | |
| Facilities | | | | | | |
| Lodging | | | | | | |
| Transport | | | | | | |

**Data Mining Workday**
**Theory and Methods**
**David Banks, Leader**
**February 4, 2004**
**Total Responses:**
**Total Participants:**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | No Formal Registration or Evaluation | | | | | |
| Staff | | | | | | |
| Facilities | | | | | | |
| Lodging | | | | | | |
| Transport | | | | | | |

**Data Mining Workday**
**Bioinformatics**
**Stan Young, Leader**
**February 11, 2004**
**Total Responses:**
**Total Participants:**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | No Formal Registration or Evaluation | | | | | |
| Staff | | | | | | |
| Facilities | | | | | | |
| Lodging | | | | | | |
| Transport | | | | | | |

**Internet Tomography and Sensor Networks**
**J.S. Marron, Leader**
**October 12-15, 2003**
**Total Responses: 19**
**Total Participants: 93**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 1 | 15 | 3 | 0 |
| Staff | 0 | 0 | 0 | 6 | 13 | 0 |
| Facilities | 0 | 0 | 0 | 10 | 9 | 0 |
| Lodging | 0 | 1 | 3 | 6 | 8 | 1 |
| Transport | 0 | 1 | 4 | 5 | 8 | 1 |

**Congestion Control and Heavy Traffic Modeling**
**J.S. Marron, Leader**
**Oct 31-Nov 1, 2003**
**Total Responses: 24**
**Total Participants: 62**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 2 | 8 | 14 | 0 |
| Staff | 0 | 0 | 0 | 4 | 20 | 0 |
| Facilities | 0 | 0 | 7 | 6 | 11 | 0 |
| Lodging | 0 | 0 | 1 | 8 | 7 | 8 |
| Transport | 0 | 0 | 2 | 8 | 7 | 7 |

**Undergrad Workshop on Data Mining**
**H.T. Banks, Leader**
**November 14-15, 2003**
**Total Responses: 29**
**Total Participants: 30**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 2 | 16 | 10 | 1 |
| Staff | 0 | 0 | 0 | 5 | 24 | 0 |
| Facilities | 0 | 0 | 1 | 10 | 18 | 0 |
| Lodging | 1 | 0 | 0 | 9 | 19 | 0 |
| Transport | 1 | 1 | 4 | 11 | 12 | 0 |

**Undergrad Workshop on Data Mining**
**H.T. Banks, Leader**
**February 13-14, 2004**
**Total Responses: 20**
**Total Participants: 21**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 0 | 8 | 12 | 0 |
| Staff | 0 | 0 | 0 | 0 | 20 | 0 |
| Facilities | 0 | 0 | 0 | 0 | 20 | 0 |
| Lodging | 0 | 0 | 1 | 3 | 15 | 1 |
| Transport | 0 | 1 | 3 | 5 | 10 | 1 |

**Multiscale Opening Workshop**
**Ralph Smith, Leader**
**January 17-20, 2004**
**Total Responses: 11**
**Total Participants: 92**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 1 | 6 | 4 | 0 |
| Staff | 0 | 0 | 1 | 2 | 8 | 0 |
| Facilities | 0 | 0 | 1 | 2 | 8 | 0 |
| Lodging | 0 | 0 | 2 | 3 | 5 | 1 |
| Transport | 0 | 0 | 3 | 4 | 3 | 1 |

**Multiscale Challenges in Soft Matter Materials**
**Greg Forest, Leader**
**February 15-17, 2004**
**Total Responses: 31**
**Total Participants: 68**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 1 | 11 | 19 | 0 |
| Staff | 0 | 0 | 0 | 6 | 25 | 0 |
| Facilities | 0 | 0 | 3 | 8 | 20 | 0 |
| Lodging | 0 | 0 | 2 | 10 | 16 | 3 |
| Transport | 0 | 1 | 4 | 8 | 14 | 4 |

**HOT TOPICS Workshop on National Security**
**Alan Karr, Leader**
**April 1-2, 2004**
**Total Responses:**
**Total Participants:**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 0 | 0 | 3 | 0 |
| Staff | 0 | 0 | 0 | 0 | 3 | 0 |
| Facilities | 0 | 0 | 0 | 0 | 3 | 0 |
| Lodging | 0 | 0 | 0 | 0 | 3 | 0 |
| Transport | 0 | 0 | 0 | 0 | 3 | 0 |

**Multiscale Workshop on Granular Flow**
**Dave Schaeffer, Leader**
**April 16-17, 2004**
**Total Responses: 19**
**Total Participants: 33**

| | 1 | 2 | 3 | 4 | 5 | N/A |
|---|---|---|---|---|---|---|
| Science | 0 | 0 | 0 | 6 | 12 | 1 |
| Staff | 0 | 0 | 1 | 1 | 17 | 0 |
| Facilities | 0 | 0 | 0 | 6 | 13 | 0 |
| Lodging | 0 | 0 | 0 | 2 | 10 | 7 |
| Transport | 0 | 0 | 0 | 1 | 11 | 7 |

## SAMSI Workshop Evaluation
## Workshop Name, Dates, 2004

Your feedback on this workshop is requested by the National Science Foundation, who view it as important for assessing and improving the performance of institutes. Your feedback is also gratefully appreciated, because it will enable us to immediately improve SAMSI workshops. Please fill this out and hand it to a SAMSI Staff Member.

**0. Personal Information:**

a. Discipline (e.g. Statistics, Applied Math.) _____

b. Highest Degree: _____ Year: _____ Current Student: ____

| **1. General Ratings:** | Poor | Fair | Good | Very Good | Excellent |
|---|---|---|---|---|---|
| a. Scientific Quality | 1 | 2 | 3 | 4 | 5 |
| b. Staff Helpfulness | 1 | 2 | 3 | 4 | 5 |
| c. Meeting Room/AV Facilities | 1 | 2 | 3 | 4 | 5 |
| d. Lodging | 1 | 2 | 3 | 4 | 5 |
| e. Local Transportation | 1 | 2 | 3 | 4 | 5 |

**2a. What were the positive aspects of the organization and running of this workshop?**

_____

_____

**2b. What parts of the organization and running need improvement?**

_____

_____

**3. Please comment on the Scientific Quality:**

a. Innovation: _____

_____

b. Communication: _____

_____

c. Level: _____

_____

**4. Additional Comments on the overall workshop / tutorial.**

_____

**5. An important goal of SAMSI is to create synergies between statistics, applied mathematics, and other disciplines. How well did this workshop further this goal?**

_____

**6. How did you learn of this workshop?**

_____

**7. Please suggest ideas / contacts for future SAMSI programs**

_____

**8. Personal Information:**

Name: _____

Affiliation: _____

Email Address: _____

The following information is for reporting purposes only. It has no bearing whatsoever on future participation in SAMSI events. Please circle appropriate choice.

Gender: Male Female

Ethnicity: Hispanic Not Hispanic

Race:

White African American Asian Native American

Hispanic Native Hawaiian/Pacific Islander Other _____

**APPENDIX G – Course Descriptions for 2003-2004**

## I. Data Mining and Machine Learning (Fall 2003)

**University Listings:**

| | |
|---|---|
| Duke University | STA 293.01 |
| NC State | ST 810T-005 |
| UNC | STAT 331 |

**Instructors:**

Professor David Banks : banks@stat.duke.edu
Professor Feng Liang : feng@stat.duke.edu

**Class Time:**

Wednesdays, 4:30 - 7:00pm
Class begins August 27, 2003

**Class Location:**

NISS Building, Room 104
Maps and Directions
Distances:

- Duke - SAMSI: ~ 8.5 miles (14 km)
- NCSU - SAMSI: ~ 16.5 miles (26 km)
- UNC - SAMSI: ~ 13.5 miles (22 km)

**Course Description:**

Data mining represents an expanding partnership between statisticians and computer scientists. This SAMSI course attempts to bring graduate students up to the research frontier in this area, drawing together the foundations (Curse of Dimensionality, smoothing, flexible modeling, recursive partitioning, and parsimony) with more recent innovations (support vector machines, boosting and bagging, model stiffness, data streaming, and false discovery rate). The class will involve some applications and some illustrative use of software, but the focus will be upon theory. Grading will be based upon a research project--students will be expected to invent a new idea in this area, implement it, and then test it (this is easier than it may sound).

**Prerequisites:** Knowledge of statistical inference, including familiarity with density functions, degrees of freedom, hypothesis testing and multiple regression; Comfort with linear models; Some experience with modern statistical computing.

**Text:** The main text for the course is Hastie, Tibshirani, and Friedman's "The Elements of Statistical Learning," but it will be supplemented by current articles.

## II. Data Statistical Analysis and Modelling of Internet Traffic Data (Fall 2003)

**University Listings:**

| | |
|---|---|
| Duke University | STA 293.02 |
| UNC-CH | STAT 321 |
| NC State | ST 810R-004 |

**Class Time:**
Tuesdays, 4:30 - 7:00pm
beginning August 26, 2003

**Class Location:**
NISS Building, Room 104

**Course Description:**
The analysis and modeling of internet traffic data represents an important major challenge for engineers, for computer scientists, for statisticians and for probabilists. Really new ideas and models are needed because heavy tailed distributions and long range dependence (both appearing at a number of different points) render standard methods, such as classical queueing theory, unusable. This course considers a variety of methods for understanding and modeling internet traffic at a variety of levels, from individual TCP traces, to monitoring traffic on a main link. An important underlying concept is cross scale views of data. Novel graphical views of data play an important role. To reach a broad audience, prerequisites are kept to a minimum, with needed foundational material, including Q-Q plots, time series analysis, long range dependence, and SiZer analysis being introduced as needed.

**Prerequisites:**
One year of probability and statistics, at the undergraduate level.

### III. Long-range Dependence and Heavy Tails (Fall 2003)

**University Listings:**

| | |
|---|---|
| Duke University | STA 293.03 |
| UNC-CH | STAT 322 |
| NC State | ST 810U-006 |

**Instructor:**
Professor Murad S. Taqqu, SAMSI, University of North Carolina & Boston University

**Class Time:**
Thursdays, 4:30 - 7:00pm
beginning September 4, 2003

**Class Location:**
NISS Building, Room 104

**Course Description:**
This course will focus on long-range dependence and heavy tails, notions which are relevant in computer traffic networks. Long-range dependence occurs when the covariances of a time series decrease slowly, like a power function. Heavy tails occur when the probability distribution of the time series has infinite variance and behaves like a power function. We will introduce self-similar processes which are idealized models that can encompass long-range dependence and/or heavy tails. We will focus first on fractional Brownian motion and on the related FARIMA time series models. To deal with infinite variance and heavy tails, we will introduce in a systematic fashion, infinite variance stable processes. We will study their properties and describe a number of stable (heavy-tailed) self-similar processes, including the so-called "Telecom model". We will also describe statistical methods for detecting the presence of long-range dependence and for

estimating its intensity, focusing on wavelet methods since these are particularly useful in this regard.

**Prerequisites:**
One year of probability and statistics, preferably at the graduate level.

**Required Text:**
"Theory and Applications of Long-range Dependence" Paul Doukhan, Georges Oppenheim and Murad S. Taqqu editors. ISBN 0-8176-4168-8. Birkhauser, Boston (2003).

**Optional Texts:**
"Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance". Gennady Samorodnitsky and Murad S. Taqqu, ISBN 0-412-05171-0, Chapman and Hall/CRC, New York (1994).

"Selfsimilar processes". Paul Embrechts and Mokoto Maejima, ISBN 0-691-096627-9 Princeton University Press (2003).

## IV. Mathematical & Experimental Modeling of Physical Processes I (Fall 2003)

**University Listings:**
NC State MA (BMA) 573

**Instructor:**
Professor H.T. Banks, Director - CRSC, NC State & SAMSI Associate Director

**Class Time:**
Mondays & Wednesdays, 2:05 - 3:20pm
beginning August 20, 2003

**Class Location:**
NC State Campus, Winston-room 104

**Course Description:**
In-depth treatment of case studies in application of mathematics and statistics to physical and biological problems arising from projects in industrial and governmental laboratories. Background information for each case study; development of mathematical and statistical models; analytical and computational methods appropriate to models; model validation using experimental data collected in the Center for Research in Scientific Computation (CRSC) Math Instructional Laboratory. Case studies involve problems in biology, thermodynamics, mechanics, electromagnetics, and hydrodynamics.

This course is offered at 2:05-3:20 MW in WN132 on the NCSU campus. It is interactively broadcast live over the NC-REN TV network to selected sites. It is part of the educational program sponsored by the Statistical and Applied Mathematical Sciences Institute (SAMSI).

**Prerequisites:**
MA 341, MA 405, knowledge of high-level programming language

## V. **Multiscale Model Development and Control Design** (Spring 2004)

**University Listings:**

| | |
|---|---|
| Duke University | STA 293.01 |
| NC State | MA 810M |
| UNC | STAT 331 |

**Instructors:**
Professor Alan Gelfand, ISDS - Duke University & SAMSI
Professor Ralph Smith, CRSC - NC State University & SAMSI

**Class Time:**
Tuesdays, 4:30 - 7:00pm
beginning January 13, 2004

**Class Location:**
NISS Building, Room 104

**Course Information:**
Introduction to deterministic and stochastic model development, numerical approximation, and control design for advanced materials. Energy-based, stochastic homogenization and statistical techniques for nonlinear model development, and full and reduced-order numerical approximation techniques for design and real-time control implementation. Statistical emphasis on stochastic modeling to accommodate scaling issues in both process mean specification and process error structure. Deterministic and stochastic techniques for robust control design for systems employing piezoceramic, magnetostrictive, shape memory alloy, and magnetorheological transducers.

**Prerequistes:**
Multivariable analysis, advanced calculus, one semester of linear algebra, one semester of numerical methods, and some previous exposure to statistics and probability.