



Statistical and Applied Mathematical Sciences Institute  
PO Box 14006, Research Triangle Park, NC 27709-4006  
Tel: 919-685-9350      FAX: 919-685-9360  
www.samsi.info

## 2003–04 Synthesis Program on Data Mining and Machine Learning

### Program Goals

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 or dimension) data sets.

The goals of the SAMSI DM&ML Program are to advance this understanding significantly, to articulate future research needs for DM&ML, 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.

Potential research foci of the program—actual foci will reflect the interests of principal participants, and will be chosen in Working Group meetings associated with the Opening Workshop—include:

- “Large  $p$ , small  $n$  inference”
- Bioinformatics
- Support vector machines
- Computational experiments that relate performance of DM&ML techniques to problem characteristics
- Text mining, and other novel, complex forms of data
- Statistical questions such as sampling, model selection and search, robustness, data quality and multiplicity inherent in many DM&ML techniques.

Computational scalability will be a pervasive theme.

### Activities

*Workshops.* There will be two major workshops:

**Tutorial and Opening Workshop**, on September 6–11, 2003 to focus the scientific agenda of the program (and also highlight the statistical importance of work being done by non-statisticians in such areas as support vector machines). Committed invited speakers include Di Cook (Iowa State), Michael Jordan (Berkeley), J. S. Marron (North Carolina), David Madigan (Rutgers) and Robert McCulloch (Chicago).

**Closing Workshop**, on May 20–22, 2004, at which findings of the program will be presented to the community.

There may also be mini-workshops (1 day) that explore ramifications of DM&ML for other issues, or treat related problems such as text mining, as well as a “mid-term” workshop in January, 2004, at which program participants will assess progress and set the course of the remainder of the program.

*Research.* Research activities, on topics such as those listed above, will be organized into *Working Groups*, each of which will focus on *testbed data sets* provided by participants or other organizations (such as the NISS Affiliates). Testbed data sets are currently being identified. Possibilities include astronomy, bioinformatics (compound activity), product performance (warranty) software engineering (change histories) and spontaneous reporting systems (example: adverse events arising from pharmaceuticals).

While research undertaken during the program is important, research and collaborations catalyzed by the program are even more important, so that formulation of the collaborations and their research agendas is a high priority.

*Monograph.* Like other SAMSI synthesis programs, that on DM&ML will produce an extensive report detailing the conclusions and recommendations of the program, which is anticipated to be published in a new SAMSI subseries of the ASA–SIAM Series on Statistics and Applied Probability.

## Opportunities to Participate

*SAMSI–University Fellows*, appointed for a semester or the entire year, will help lead the research, as well as join academic life at a SAMSI partner university; see [www.samsi.info/univfellows200210.html](http://www.samsi.info/univfellows200210.html) for further information. Salary and timing are ideal for senior sabbatical visitors with partial funding from their home institutions.

*Long-term Research Visitors* will participate in the research; support for expenses is available.

*Postdoctoral Fellows* can be appointed for two or more years; see [www.samsi.info/postdoc200210.html](http://www.samsi.info/postdoc200210.html) for information and application instructions.

*Workshop Attendees* both inform the course of the research and have early access to the results, and may receive support for expenses. Workshops will be announced individually on the SAMSI Web site.

For further information, write to [dmmml@samsi.info](mailto:dmmml@samsi.info). Members of underrepresented groups are especially encouraged to apply.

## Scientific Committee

David Banks (FDA; Co-chair), Mary Ellen Bock (Purdue), Jerome Friedman (Stanford), Alan F. Karr (NISS; Chair), David Madigan (Rutgers), William DuMouchel (AT&T), Warren Sarle (SAS Institute).

## About SAMSI

The Statistical and Applied Mathematical Sciences Institute (SAMSI) is a national institute whose vision is to forge a new synthesis of the statistical sciences and the applied mathematical sciences with disciplinary science to confront the very hardest and most important data- and model-driven scientific challenges. SAMSI achieves profound impact on both research and people by bringing together researchers who would not otherwise interact, and focusing the people, intellectual power and resources necessary for simultaneous advances in the statistical sciences and applied mathematical sciences that lead to ultimate resolution of the scientific challenges.

SAMSI is a partnership of Duke University, North Carolina State University (NCSU), the University of North Carolina at Chapel Hill (UNC), and the National Institute of Statistical Sciences (NISS), in cooperation with the Mathematical Sciences Research Institutes program of the Division of Mathematical Sciences at the National Science Foundation and in collaboration with the William R. Kenan, Jr. Institute for Engineering, Technology and Science. SAMSI is located at the NISS building in Research Triangle Park, North Carolina.