

## Summary

### The IPAM short program on Mathematical Challenges in Scientific Data Mining

The IPAM short program on Mathematical Challenges in Scientific Data Mining was held during January 14-18, 2002. The goal of the program was to bring together mathematicians, data mining practitioners, and domain scientists in order to share experiences and identify the open mathematical challenges that must be addressed for data mining to be successfully applied to the analysis of data in a variety of scientific disciplines.

Data mining is the discovery of patterns, associations, anomalies, and statistically significant structures in data. It borrows and enhances ideas from several different fields, including signal and image processing, statistics, image understanding, mathematical optimization, computer vision, high performance computing, and pattern recognition. Data mining techniques can be applied to the analysis of large and complex data sets in several scientific domains such as astronomy, remote sensing, bio-informatics, medical imaging, non-destructive evaluation, combinatorial chemistry etc. While these domains provide a rich and challenging environment for the practice of data mining, they are often overlooked in data mining conferences, where the focus tends to be on commercial and business data. As a result, this IPAM program provided a wonderful opportunity for interaction among researchers focusing on problems specific to the mining of scientific data sets.

The program, which took almost 9 months to plan, consisted of four and a half days of presentations. Given the variety of topics that comprise scientific data mining, and well as the diverse backgrounds of the participants, the program was organized as follows:

**Tutorials:** The program started with two tutorials. The first was a survey of applications of data mining in science and engineering. It also included the processing of the raw data in the form of images and meshes to prepare it for the identification of patterns in the data. The second tutorial was on the pattern recognition algorithms. The goals of these tutorials were not only to provide an introduction to the subject, but also to provide the context for the topics covered during the week.

**Introductory talks:** There were 11 introductory talks, each lasting an hour. These covered both the algorithms and the applications aspect of scientific data mining. The former included algorithms for data preprocessing such as wavelets and independent component analysis, as well as algorithms for pattern recognition such as ensembles of classifiers and support vector machines. The talks on application of data mining covered areas such as computer vision, and the analysis of data from micro arrays, brain images, and earth sciences. These talks, especially those on applications, typically included an introductory part to familiarize the participants with the problem domain, and concluded with the speaker's opinion on the mathematical challenges in scientific data mining.

**Specific talks:** There were 11 such talks covering a specific topic in greater detail. Each lasted 45 minutes. The talks on algorithms covered topics such as the use of level sets and partial differential equations in image processing, mixture models, and graph theoretic

approaches to finding frequent patterns. There were also talks on the application of data mining to specific problems such as protein structure prediction, mining of biological sequences, mining solar imagery, and mining the global climate system.

**Contributed talks and posters:** In addition to the invited talks, the program also included contributed presentations that resulted from an open call posted to several data-mining mailing lists. The abstracts submitted were reviewed by the organizers for their suitability to the program theme. All but one of the submissions were accepted. Of these, 15 were presented as posters and 5 as 30 minute talks. The poster session, which overlapped with an evening reception, was very well attended, and several participants stayed long after the close of the session. There was also an impromptu poster contribution from one of the participants resulting from discussions during the week. Several of the posters were by students and recent graduates.

**Special session on data mining system architectures:** Since several of the participants had developed end-to-end data mining systems, a special session was also held in which the speakers discussed some of the non-mathematical issues encountered in applying data mining to scientific data. These included the handling of different data formats, the ability to try out different algorithms for a problem to find the most appropriate one, and the ability to iteratively refine the process until desirable results were obtained.

The program was very well attended, with just over 100 registered attendees and several local attendees who attended talks of interest. The program also supported full or partial travel and lodging for 12 students and post-docs. The audience included scientists from Europe, Japan, and the US. Many of the participants stayed the entire week, attending all the talks. The presentations led to active discussions during the question/answer period as well as during the breaks in the program. The cultural barriers resulting from the diverse backgrounds of the participants dissolved quickly as common threads were identified and solutions to problems were found in fields very different from one's own.

The response to the program was overwhelmingly positive. Almost all the participants requested a sequel to the program, with several of the requests made on the very first day! Many participants commended the diversity and relevance of topics as well as the high quality of the talks and posters. As the lead organizer, I received several compliments both during and after the program; this credit is to be shared with the IPAM staff who provided such a wonderful environment for a productive meeting and the IPAM directors for giving us the technical freedom to explore challenging problems in this exciting and rapidly evolving field. The IPAM building itself was very conducive to promoting discussion and helping to set up collaborations among the participants. It is my sincere hope that IPAM will be amenable to hosting a follow-on meeting where we can further explore and understand the mathematical challenges in mining scientific datasets.

Chandrika Kamath  
Center for Applied Scientific Computing  
Lawrence Livermore National Laboratory  
February 15, 2002.