Editorial

Usama Fayyad
Data Mining & Exploration (DMX) Group
Microsoft Research
Redmond, WA 98052-6399, USA
Fayyad@acm.org

With the beginning of the year 2000, and according to some the beginning of the second millennium, we are pleased to publish the second issue of *SIGKDD Explorations.* I have toyed with the idea of calling it *volume 1, issue Y2K,* but since the Y2K term will live to be a hallmark example of over-hype and no consequence, I thought it better to avoid it and stick with the factual *volume 1, issue 2.* Instead, I decided to share with you the Dilbert comic strip below, which should serve as a warning to us all to watch out for hype, traps, and the *mines* of data mining (of the explosive and harmful variety). But now let us move on to more serious matters.

I would like to take this opportunity to welcome **Sunita Sarawagi** to the editorial staff of the newsletter. Sunita joins SIGKDD as ***Associate Editor*** of ***SIGKDD Explorations***. She has played an instrumental role on this issue, and will continue to take increasingly important roles on future issues of this newsletter. Some of you may know Sunita from her publications in database and data mining conferences and journals. She has worked for a few years at IBM Almaden Research Center and moved on to become a faculty member at the Indian Institute of Technology (IIT Bombay). I am confident that her qualifications, energy, and knowledge of the field will be a huge asset to *Explorations* and to SIGKDD. Welcome aboard Sunita!

This issue of *Explorations* continues to bring, we hope, timely and informative articles on topics related to the field. We have included many KDD-99 conference reports since this is the issue that follows our annual conference. Overall, KDD-99, the fifth in the series, and the first to be held under the auspices of ACM and SIGKDD was very successful. A conference report by the chairs is included in this issue and gives an overview of the state of KDD-99. Other reports cover KDD-99 workshops and results from the KDD Cup competitions. We have also included reports from other related conferences and workshops.

This issue also includes survey articles, position papers on (hopefully) controversial topics, and of course contributed technical articles. As in Issue 1-1, we have included a book review.

In future issues of *Explorations*, we hope to evolve this newsletter into a special topics format where each issue includes a collection of papers focused on a narrow topic of interest. Suggestions for special topics, and guest-editor volunteers for this purpose are always welcome. Please e-mail such suggestions to me or to Sunita (sunita@it.iitb.ernet.in).

Remarks on the Field

Data Mining and Knowledge Discovery has indeed come a long way as a field of academic study as well as a field of commercial practice. I would like to take the opportunity to comment briefly on both aspects.

Data Mining on the Academic Front

While the formation of ACM SIGKDD and its success so far is an important organizational development, I would like to focus on the purely academic/technical front. In my opinion, 1999 has witnessed the field passing through two major important academic milestones. The first is the dramatic rise in quality of the papers submitted, papers accepted, and reviews of papers in KDD-99. Credit goes to Surajit Chaudhuri and David Madigan for increasing the length of submissions to 20 pages, for insisting on thorough reviews, and for setting the most stringent of standards on acceptance. An acceptance rate of about 10% puts KDD-99 research papers at a more selective peer-review filtering level than all conferences I am aware of, indeed more selective than most technical journals in the field.

The second important academic milestone related to the field is the recent ranking of journals by *Journal Citation Reports -- Science Edition*. Scott Delman, the Publisher at Kluwer of *Data*

DILBERT reprinted by permission of United Feature Syndicate, Inc. [ref. 12727]

Mining and Knowledge Discovery, has relayed to me the following information regarding our principal technical journal in the field (in what follows, I am mostly quoting Scott):

"The JCR, published by the Institute for Scientific Information ranks journals by impact factor, which compares the average number of article citations a journal receives in a given year against the total number of articles published in that year, so this ranking is not subjective.

In the Computer Science-Information Systems rankings, *Data Mining and Knowledge Discovery* ranked #11. In the AI category in 1998, *Data Mining and Knowledge Discovery* had an impact factor of 1.235. It is highly unusual and an extremely good sign that such a new journal (less than 100 articles published so far) would even make it into the top 20 journals in the JCR rankings. The journal is ranked well ahead of many well-established, recognized, and well-respected journals in these two categories."

Data Mining on the Commercial Front

There is a veritable explosion of commercial activity around the area of KDD and data mining. The activities span a healthy spectrum from consulting and services, to products, to embedded applications, to government and scientific applications. The range of companies is also varied including major established players all the way through medium, small, and startup companies. The key drivers remain to be the explosive growth in data generating, data gathering, and data storage devices. Data seems to always expand to fill whatever container available to hold it. Hence, I foresee no sign of an end to the data avalanche. Data overload is simply becoming a way of life for companies and individuals, and as we have done throughout human history, we will build ways to better control and exploit the environment we're in. Data mining is a key technological piece in this activity.

There is also a maturation process on the commercial front. Companies are realizing that *convenience* and *integration* are key factors in making a technology useful. Database vendors are embedding data mining extensions, and data mining developers are beginning to use such capabilities to build embedded data mining applications on top.

However, not all is well in Data Mining Land. With the maturation and adoption of a technology, comes the need to clearly and simply articulate the benefits and measure the rewards. This is where we, as a field, are not doing well yet. There are no clear, agreed upon, metrics for measuring returns from data mining applications. For example, in marketing applications, data mining tools and companies have not attempted to standardize measures of return and measures of cost. There are no clear solid standards of practice. It is still a business model which is based on convincing a prospective client to take chance, spend a huge budget, and just hope that somehow, *magically*, business will get better. This mode needs to change to one where the process is more clear and less risky. I am aware of many instances where data mining has added tremendous value and has solved significant problems. However, these successes have been limited in scope and did not involve an end-to-end complete cycle. I have no doubt that successes will continue and methodologies will evolve. I simply would like to point out that we should take note of what's missing in the picture, and not focus only on the positive developments.

Concluding Thoughts

At KDD-99, there was much talk about whether KDD is a technology that is meant for widespread adoption. In his invited talk, Rakesh Agrawal wondered whether we would "cross the chasm", the divide between technologies that get early adopter enthusiasm, and those that become mainstream and widely adopted. Technologies that do not cross "the chasm" are doomed to stay small or disappear. This talk generated much discussion and introspection among attendees and at the conference and afterwards (which is exactly what a good invited talk should do).

However, I'd like to take this opportunity to throw in my view (taking unfair advantage of this unchallenged forum) that this question is not really fundamental to our field. I do not believe we should aim to be widely adopted (like television, radio, telephone, calculator, etc). Rather, we need to target making sure we develop the techniques to solve the narrow tasks of data reduction and visualization. Data mining can only succeed by being invisible and integrated in other technologies (such as database systems and vertical applications). The field is about algorithms, techniques, and methods, not hardware gadgets. Also, I do not see our anticipated user base to extend beyond use in the organization. As George John argues very eloquently in his article in Issue 1-1 of *Explorations*, data mining probably affects all of us in ways we are unaware of already. However, those who use data mining techniques directly will, by definition, be always very few in number. This is much like the number of users of programming languages, or of marketing tools, or special purpose devices that enable bigger capabilities (such as database server, network switches, etc.)

So don't be discouraged if in twenty years your uncle or your grandchildren are unaware of data mining. I only hope that we'd still be around to report about it and contribute to its advance as a field.

About the editor:

Usama Fayyad is a Senior Researcher at Microsoft Research (http://research.microsoft.com/~fayyad) where he heads the Data Mining & Exploration (DMX) Group. His work with Microsoft product groups includes the development of mining components that ship with Microsoft Site Server (Commerce Server 3.0 and 4.0) and on developing scalable algorithms for mining large databases and architecting their fit with server products such as Microsoft SQL Server 2000 and OLAP Services. After receiving the Ph.D. degree from The University of Michigan, Ann Arbor in 1991, he joined the Jet Propulsion Laboratory (JPL), California Institute of Technology, where (until 1996) he headed the Machine Learning Systems Group and developed data mining systems for automated science data analysis. He is also an adjunct professor of computer science at USC. He was also affiliated with JPL as a Distinguished Visiting Scientist. He is a co-editor of *Advances in Knowledge Discovery and Data Mining* (MIT Press, 1996) and is an Editor-in-Chief of the journal: *Data Mining and Knowledge Discovery*. He served as program co-chair of KDD-94 and KDD-95 and as general chair of KDD-96 and KDD-99.