Editorial

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

From a *Computer World* Interview, by Gary Anthes (January 10, 1999)[1] with Nobel laureate, and former chief of Bell Labs: Arno Penzias

CW: What will be the killer apps in the corporation?

PENZIAS: Data mining will become much more important. Your bank will know everything you've bought. Companies will throw away nothing they know about their customers because it will be so valuable. **If you're not doing this, you're out of business.**

Could I think of a more optimistic quote to start this very first issue of the newly founded *SIGKDD Explorations*? Not likely. Indeed, those of us who strongly believe in the growing importance of this area and the increasing need for data reduction and analysis tools are hardly surprised. Data collections continue to grow, seemingly out of control. Most organizations are doing little to extract useful knowledge from the growing mountains of data. There is a huge opportunity for data mining and other data reduction techniques to add tremendous value. This is the opportunity for people working in our new but growing field. Given the competitive pressures on organizations to become more efficient and more responsive, it is not difficult to see that much remains to be achieved. It is our hope that SIGKDD will make it possible to advance the research and applications in data mining and knowledge discovery in databases faster and in a more focused fashion.

A New Phase in Growth: KDD turns 10.

The launch of the new ACM SIGKDD signals a new stage of growth and development for the growing and dynamic field of Data Mining and KDD. Ten years ago, a mysterious invitation letter appeared in my (then fairly manageable) e-mail inbox. Gregory Piatetsky-Shapiro, then a researcher at GTE Laboratories in Waltham, MA, wanted to hold a workshop about "Knowledge Discovery in Databases". The name KDD sounded very intriguing back then. A few groups, scattered around the globe, were exploring possibilities of applying pattern recognition, machine learning, statistics, AI, visualization, and to some limited extent database techniques to the problem of making it easier to extract meaning out of data. The various groups used different names, came from multiple fields, and used a variety of appellations to describe what they were doing. The majority were Machine Learning and Statistical Pattern Recognition types. But there was a clear commonality of objectives that drew them to downtown Detroit, MI to attend the first workshop on KDD. Back then I was still a graduate student, spending a summer internship at JPL. I felt I had to attend this exciting meeting. I was not at all disappointed. Most of us knew back then that more exciting work and many more larger meetings lay ahead.

While for as long as people kept and stored data, they understood that data could hold huge value, the groups that assembled at that first KDD workshop had a distinctly computationally-intensive approach to achieve that goal. They all seemed to have at least two things in common:

1. a belief that tools for data reduction and abstraction are a must if we are to exploit our growing data assets.
2. a belief that algorithms and programs can do much of the work of reduction and in some cases discovery (i.e. constructing new models and scoring them) in a mostly automated fashion.

The words KDD struck a strong chord with many invitees. Some 60 attendees showed up at the first KDD workshop. The first few bi-annual workshops witnessed a tremendous growth, culminating in the first international conference at Montreal, Canada: KDD-95. In the meantime, KDD and data mining work in the database community was growing dramatically. The attendance at the KDD conference kept growing, reaching some 750 attendees at KDD-98. In the meantime, many developments have occurred in the past decade: databases have grown in adoption and have gained a critical role in all sorts of enterprises, our ability to gather data has grown even faster, the Web became ubiquitous and web servers themselves are now generating a flood of data, and my e-mail inbox (as I am sure many of the readers' inboxes) have grown past the manageable: bringing the KDD problem extremely close to the personal level.

A clear need emerged for a more formal organization to address the issues and to manage the growth. For this reason SIGKDD was formed as a special interest group of the ACM. There are important milestones that mark the beginnings and growth of any new field. I consider these events to be: the successful formation and growth of our annual international conference (KDD series), The first two collection volumes on KDD (1990 and 1996), and the successful launch of our journal (*Data Mining and Knowledge Discovery)* in 1997. The successful launch of SIGKDD itself is a culmination of these milestones. What better way to celebrate our tenth anniversary? I sincerely hope the field continues on this healthy path of growth and improvement.

Won Kim, the SIGKDD chair, introduces this new group in his letter in this issue. In this editorial, I would like to focus on the rationale behind the newsletter and on some general comments about the field in general.

[1] To read the full text of the interview, see Computer World's site: http://www.computerworld.com/home/print.nsf/all/990104211A

Why Explorations?

This newsletter is intended to provide a rapid mechanism for publishing work, opinions, statements, and news of general interest to the community. The intent is to let *Explorations* grow in whatever direction the community takes it. The goals are to publish relevant technical articles of interest, position papers, book reviews, general comments, and relevant news and events. It is also a venue for summarizing events of interest at the KDD conference (panel, workshop, and session reports are very welcome indeed). Reviews and summaries of happenings of interest at other related conferences are also very welcome. Position papers of interest to the field would fit in very nicely. Papers are not reviewed, except by the Editor (or Guest Editors), for style and relevance. Technical accuracy and soundness are left to the authors' judgment. I see *Explorations* as slowly evolving to a format where each issue will be a special issue focusing on an area of interest within Data Mining and KDD. Papers are not intended to be technically deep, nor does publication in this newsletter constitute a peer-reviewed publication. The intent is to provide a written record for important developments in the field, and to cover ideas, events, and issues of interest to researchers and practitioners in the field.

In this Issue

This issue includes several examples of the types of articles we would like to publish in the future. We have conference reports from KDD-98. George John reports on the popular panel on when data mining will become a successful technology. Ismail Parsa reports on the extremely well attended industrial exhibits session. The popularity of the talks associated with the exhibits gave birth to a formal track in KDD-99 focusing on industrial experience presentations. John Elder and Arnold Goodman report from a closely related conference in the Statistics community: the Interface-98 conference. This conference continues to emphasize data mining. Slowly but surely, more data mining sessions are beginning to appear in large conferences in statistics.

Review and survey articles include some words by David Hand on the relation between KDD and Statistics. Statisticians have been slow to realize the crucial importance of statistics to KDD. However, they are now paying increasing attention to this area. A survey of tools by Michael Goebel and Le Gruenwald reviews some of the tools that are commercially available. Because KDD is an extremely rapidly growing field, a comparative survey is bound to be dated. However, the methodology of comparison should be very instructive and interesting. A bibliographic summary of temporal, spatial, and spatio-temporal data mining research is covered by J.F. Roddick and M. Spiliopoulou. Finally, an application article is given by Gruca, Klemz, and Petersen. In what will hopefully be a growing section, we have a book review by Karim Hirji. Several news and announcements submitted are covered in the last two sections.

In future issues, I would like to see more position papers (as in David Hand's paper) and more application-oriented summaries. We need to document both success and failure case studies in KDD—articles along that front are very welcome. Finally, more review articles of workshops, panels, and events at the KDD and related conference should be a target of *SIGKDD Explorations.*

From Introspection to Making it Happen

The first few years of KDD, much like any other nascent field, witnessed a great deal of questions and debates about the name, identity, and what the core problems of the field are. Initially, researchers from multiple fields thought they could address data mining in isolation within their respective fields. Database conferences continue to have many sessions on data mining, but the papers are notably lacking in their statistical content. Statistics conferences have many more sessions on data mining, but the word "database" hardly appears anywhere in those papers. The same can be said of sessions and conferences in many areas including: AI, machine learning, data visualization, Information Retrieval, optimization and operations research, high-performance computing, and neural networks. The KDD conference and the journal *Data Mining and Knowledge Discovery* (http://research.microsoft.com/datamine) remain the best places where balanced coverage of work that involves contributions from many related fields appears. Slowly but surely, work in data mining is taking on its own flavor as a distinct area of research.

The most encouraging signs are less pre-occupation with what the field is all about, and much more attention paid to developing new algorithms, implementing scalable systems, and solving new application problems. We have now entered the long and arduous stage of building sufficient infrastructure and laying mathematical foundations to what we do. Progress in this stage will be slow, but systematic. Developments will not be as exciting, but their importance will be greater. We have moved from asking questions like: *who are we? What do we call ourselves?* and *what are the problems we solve?* to asking questions of the following sort: *what are the systems issues? What does a DBMS need to support? What are the proper abstractions? What should be the standards? What are fair and meaningful benchmarks?* and so forth.

In addition to the continued growth of the KDD academic community, there is now a healthy number of commercial activities in the area. The market for data mining is growing. Corporations are beginning to demand more integrated solutions, and practitioners are quickly realizing that systems issues can be become overwhelming. This is only the beginning of a long journey. To make it through, we need to get systematic about the problems we address, the solutions we deploy, and understanding the limitations of our existing technology.

The Market is Confused: Who is the Customer?

Both in industry and in academia we need to be clear about the market niche we serve and the set of academic and scientific problems we address. While at the technical level the scope of problems we address is very deep, it is important to understand that the final goal is to make the techniques useful to users. Better algorithms for clustering in high dimensions, for density estimation over large non-homogenous databases, or for discovering sequential patterns necessarily require formal mathematical approaches and fairly counter-intuitive algorithms and methods. However, at the end of the day, someone needs to be able to understand what the techniques deliver. Without an easy way to visualize the model it extracts, the most advanced algorithm can remain unused for a long time (and perhaps forever or until rediscovered again). It is important to insure that the inputs and outputs are made simple to understand.

Looking at today's market, I see a large number of vendors selling data mining tools. However, few of these vendors are really offering anything beyond yet another statistical library. While I can understand the need for new algorithms (perhaps more scalable, more autonomous, etc), the problem is these tools seem to be targeted at "end-users". By "end-user" I mean someone whose focus is not primarily data analysis, but actually a business function (marketing manager, banking officer, knowledge worker, etc.) It strikes me as rather strange to think that a business user would ever be interested in data mining technology. The end-user is interested in *solutions* and not in technology. No wonder some of the data mining companies are having trouble "selling" their wares. Unfortunately, faced with the threat of market failure because their products are targeted at the wrong customer, some vendors resort to hyperbole and false promises. It is my hope that the market will mature quickly and learn to ignore the hype. Fortunately, there are good signs of this already.

In my opinion, the target should be to make it easy for *solution providers* to deploy solutions that utilize data mining technology. Much like a database system contains a lot of algorithms in its I/O system, its query optimizer, and its storage engine, we do not expect to "sell" such features to end-users. They simply enhance performance. Similarly, goals of data mining algorithms should be to make it easy for a software engineer (someone who builds solutions) to put together a good data reduction, prediction, or reporting utility. The technology needs to be embedded and standardized for it to achieve wide adoption and impact. Telecommunication and transportation technologies have achieved this. To a large degree computer networking technology has achieved this: *I would not recognize a packet-switched network if it hit me in the face; but I use it daily.* We do not expect a network user to understand what happens at the various layers or what routing protocol even means.

The key is to create a standardized simple programming interface, and to try to hide much of the details whenever possible. Methods should take care of their own thresholds, parameters, and other standard hygiene functions like their accuracy settings and optimization. Most respectable models or patterns will tell a sufficient story on what's happening in the data. Focusing on the micro-details of improving the accuracy of an algorithm is an internal matter that should not be left to the user to decide. Hence our target customer is really an engineer who understands software and has some exposure to the target application.

The other engineering challenge is scalability to large databases. In the recent few years, this latter challenge has received much attention, at least at the level of dealing with a large number of data items. However, dealing with high dimensionality is still an important scalability challenge. I see too little research on making methods simple to use, on visualization techniques, and on understanding how to trade off complexity and accuracy.

The Challenges Ahead

There are many challenges ahead of us. It is rather difficult to correctly guess where the real challenges will be. However, I shall venture a few guesses despite the fact that there is a good probability that I am mistaken. Most importantly, especially in our research, we need to avoid the hyperbole. Rather than claiming we are working on "intelligent" algorithms that "learn" from data, we need to recognize that data mining is about data reduction and modeling. We need to understand that integration with database systems is necessary (because that's where the data are). Use of data mining is predicated on the convenience of using it. It is also predicated on clearly understanding its limitations. This involves educating the target customers and the technical community about the pitfalls. Data mining models are not deterministic entities. The process is sensitive to data quality, data sample size, and other estimation hazards. The results are never "correct". The operations are inherently *probabilistic*. So once in a while, an mining algorithm will extract garbage instead of gold. This is a fact of life that users can tolerate if the frequency of "garbage" is kept low.

We need to define the standards and understand what the proper abstractions are. What does one need to do mining? What does a system need to support? Where is the real cost in doing data mining? Many other important questions need to be investigated.

Why is integration with a DBMS so important? Practitioners understand that most of the cost of "doing' data mining is not in the modeling algorithms. Rather it is in *data cleaning and preparation*, in *model deployment*, and in *maintenance and management.* Providing a standard API and integration with a database environment can reduce much of the cost associated with the latter two aspects. Data preparation, selection, and cleaning remains a looming challenge that faces databases, data warehousing, and decision support systems in general.

Another important challenge is to establish and document success stories. Because of the competitive nature of businesses, much of the details of great successes of data mining remain unpublished as customers view them as trade secrets and competitive advantage matters. We need to address this serious problem. Success stories are one of the most important ways for a technology to gain adoption. This will be a tricky challenge for our field. I remain hopeful that a good solution will emerge.

About the editor:

Usama Fayyad is a Senior Researcher at Microsoft Research (http://research.microsoft.com/~fayyad) in 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 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 received the 1994 NASA Exceptional Achievement Medal and the JPL 1993 Lew Allen Award for Excellence in Research for his work on developing data mining systems to solve challenging science analysis problems in astronomy and remote sensing. 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.