Editorial

Usama M. Fayyad
DMX Group
601 108th Avenue NE, 19th Floor
Bellevue, WA 98004, USA
Fayyad .at. DMXgroup.com

Some Great News

Some time has elapsed since I last addressed the readers of *SIGKDD Explorations* from this forum. It is with particularly great pride and excitement that I do so this time. Your newsletter has undergone a change in primary editorial leadership and is moving to the next stage on its evolution to being an established and well-read publication.

With this issue *Explorations*, the role of Editor-in-Chief has been fully transitioned to Sunita Sarawagi. In addition, much of the primary associate editor work is now solidly in the hands of Paul Bradley.

Both Sunita and Paul have been doing an outstanding job with *Explorations* for the past two years as Associate Editors. In fact, they have taken much of the load off me – so much so that I have done very little in the past year. This is as it should be when very energetic and motivated leadership joins an organization. Sunita and Paul have been exemplary associate editors and have proven that they deserve my trust and more importantly the trust of the SIGKDD Executive Committee. I am certain also that through their performance over the last two yeas, they have gained the trust of the SIGKDD membership and the community at large.

At KDD-2003 in Washington, D.C. we announced that Sunita would take over the Editor-in-Chief role. Sunita also addressed the community during the SIGKDD business meeting and briefly described some of her plans for the newsletter. I believe she and Paul will continue to provide excellent leadership and will take this newsletter to new heights. It is my hope that the membership and the community will work with them in making *Explorations* a success and in helping it to advance the sense of community and scholarship in this field. Stay tuned for further developments and additions to the editorial team over the next few months as I am sure Sunita will be updating you on these developments. From my side, I would like to express my deep gratitude and admiration for their exemplary service over the past 2 years.

The Research Community is Growing

With the close of the year 2003, we close our fifth year of SIGKDD as an ACM organization. The KDD conference is now in its ninth year as an international conference, and in its 15th year since the first KDD-89 Workshop, with KDD-03 being the 13th meeting in the series. The conference has proven to be a steady and healthy conference, with attendance exceeding 700 people. More importantly, it has maintained a high quality track record for research papers, workshops, tutorials, and events such as the KDD Cup. It is an extremely competitive forum for publication, with less than a 10% acceptance rate for research papers. It is more difficult to publish a research paper in KDD than most other conference, and certainly more difficult than publishing in the majority of technical journals. The KDD conference has continued to grow and improve in quality despite the proliferation of alternative conferences in the U.S. and internationally. The emergence of other major conferences is a very healthy sign for the field in my opinion.

On the other front, the journal: *Data Mining and Knowledge Discovery* has also continued to grow. Now in its eighth year in publication, the journal has recently expanded from 4 issues per year to 6: Another healthy growth sign for the community. Other activities by the SIGKDD EC and its subcommittees, especially the Curriculum Committee chaired by Jiawei Han are further signs of healthy and disciplined growth on the academic front. More encouraging to me is the observation that most top universities now have established formal tracks or curricula focused on KDD and Data Mining.

The Forces of Nature

With the ease of capturing and storing data, large databases and massive data sets and warehouses have become the norm. aradoxically, despite the proliferation of large data stores, it is a rare event to find an organization that is actually benefiting from its data assets. About 5 years ago I wrote that most large data warehousing projects appear to be failing to deliver value to their owners, and I called them “write-only” stores. Unfortunately, this fact has not changed much since that writing; in fact the situation has been has been getting worse by the month. Sadly, the analogy of the typical data store to a *data tomb*, where data is deposited to rest undisturbed and unused remains very appropriate today.

We continue to exist in an environment that is driving the need for the kinds of technologies and applications that the KDD field is poised to address. Our fundamental problems are ones that are becoming increasingly important to society, by the day, because of the ever-growing cache of data in each organization. With the continued proliferation of the digitization of transactions, workflows, and processes: be it in science, business, government, communications, or security, one axiom appears to continue to hold true: growth in stored digital data seems to be keeping a surprisingly fast pace – a pace whose trend eclipses Moore’s law.

While these large untapped resources present a huge opportunity for our field, I would be very concerned if we, as a community, do not contribute in a significant way to impact this environment in a major way over the next few years. While, admittedly, progress in research and science is typically slow, we need to ensure that at least some of the advances we have been developing in the KDD field start gaining serious deployments in practical systems.

I see this need for true deployments of the new technology as the prime challenge for our community to make itself relevant.

Another challenge is to continue to attract contributions and advances from researchers and practitioners in related fields, including statistics, databases, optimization, AI and machine learning, and visualization. There are many relevant areas in applications. These are too many to list but include: bioinformatics, genomics, physics, many disciplines in science data analysis, as well as computer/network security and systems. Without gaining the attention and involvement of people from these related communities, we all stand to lose much.

In my opinion, one key to attracting involvement of other communities is to have a clearly elucidated set of contributions and well-understood challenge problems. I believe we face many challenges on two primary fronts: the pragmatic (or practical) front as well as on the technical (or theoretical) front. Work in KDD and data mining has been taking place along so many areas and on so many fields that it is difficult to assess progress in a coherent way. While having much activity without a coherent centralized theme is perfectly acceptable and natural in a young field like ours, I believe that this should not prevent us from having the basic framework for assessing progress along the fronts that we deem particularly important.

Initial Proposal of Grand Challenge Problems

I propose formulating a set of Grand Challenges that over time become important markers by which we are able to assess our progress. I take an initial attempt at such a list below, and I hope that over time we refine this list through input from all who are interested. The goal is to evolve a clear set of easy to state Grand Challenges that serve to attract new researchers as well as provide us a sense of progress and achievement over the long range. While such an activity is difficult, and some might say inappropriate for a young field, as long as we keep this a living, active and changing list, and as long as we are honest in assessing progress (or the lack of it) along these challenges, I think such a list will serve some very useful purposes for the community. In addition, I believe that such a list of Grand Challenges will serve as a beacon for interested contributors from other communities.

My initial list of Grand Challenge problems is divided into the Pragmatic and the Theoretical. It is my sincere hope to collect feedback from the community and revise this list over time until we reach such stage where a majority of the community believes these problems are worthy and sufficiently representative…

To keep matters simple for this initial stab at establishing grand challenges, I would like to keep the discussion general. Furthermore, while I have many challenges in mind, I will restrict myself to a total of 10 challenges, five within each category.

Technical Grand Challenges

1. How does the data grow? We need a theory for fundamentally understanding what "large" or "massive" really means. I first heard this problem from Peter Huber in 1997, and I have yet to see it addressed. Large databases never grow as a result of an independent sampling from a static distribution. This IID assumption permeates much of the literature. We inherited it from statistics and engineering. We are in regimes where clearly large databases grow over time from dynamic and changing sources. Yet we don't possess the elementary tools for modeling this situation and describing it formally.

2. Complexity/Understandability Tradeoff. We need the mechanisms and techniques for taking complex models that optimize a metric like predictive accuracy, and render them understandable. In the reverse direction, we also need to be able to start with a simple strategy statement and generate a complex model from it. I often say that *An Ounce of Knowledge is Worth a Ton of Data.* We need to be able to leverage general knowledge of the problem to form complex and detailed models.

3. Interestingness. Much work has considered measuring when a pattern is interesting, useful, or novel. Yet we still have no theory of how to go about this fundamental notion in the field.

4. Scalability. Making our algorithms not only work with large volumes of data effectively, but high dimensionality, variety of data types (text, images, audio, structured, unstructured, etc.) This includes integration with database management systems. It also includes defining a framework for graceful degradation between high-dimensionality and high fidelity in reduced subspaces -- the equivalent of the SVD (singular value decomposition) in our field.

5. A Theory for What We Do. What are the fundamental abstractions? What are the basics operations? What are the basic components of an algorithm? What is it that we are optimizing? What is hard? What is doable? Why? What is a "data summary"? When are two attributes "similar" or "partially similar"? Can you measure this partial similarity efficiently? Many other questions under this heading including Heikki Mannila's favorite (and I paraphrase): what is an algebra for dealing effectively with highly condensed (or summarized) views of large data universes? There is a much longer list here, but any advances along the theory for what we do in KDD should be a significant contribution.

Pragmatic Grand Challenges

1. Where is the Data? While we speak much of the data overload, in most situations when one starts a data mining project in an organization or enterprise, data somehow becomes scarce or inaccessible. We cannot simply wait on corporate cultures to change or for people to figure out how to solve the chronic problems in the field of data warehousing. We need to let our algorithms go where the data are, and we need to build our solutions so they contain the necessary data management infra-structure (while leveraging existing tools and systems, most notably database technology ad systems). It is imperative that data mining algorithms not be starved for data, and it is up to us to solve this "shortage" rather than wait for it to be solved for us. If we do, I predict a long and fruitless wait.

2. Embedding algorithms and solutions within operational systems. The time is now for data mining researchers and practitioners to "get our hands dirty". While we all prefer to specialize and focus on what we like, the nature of our field demands a serious degree of immersion and involvement in the surrounding context and operational systems. For an algorithm to truly be interesting, it needs to solve a real problem and be robust to all the variety and inconsistency that comes with every day life.

3. Integrating Domain Knowledge. We all know that data mining algorithms are "knowledge-free", meaning they are brittle and in real applications lak even the very basic "common sense reasoning" needed to recover even from simple situations. We

also all know how difficult (some might say impossible) is the problem of common sense reasoning. There is a surprisingly effective means to solve this problem: highly focused and deeply integrated solutions. When a data mining algorithm is designed to solve a very specific problem, sufficient knowledge can be embedded within the algorithm and its harness, that it will appear to exhibit deep common sense reasoning ability and it will suddenly become robust to all sort of anomalies. I have yet to see an algorithm in data mining that integrates such domain knowledge.

4. Managing and Maintaining Models. While our community is good at producing new algorithms with every conference, meting, application, or occasion, I see few if any that consider the problem of model maintenance and management. What happens with historical models? How are they retired? When are they updated? How are the updates performed so the "world" is not changed on the users or the entities being serviced by these models? These are important issues if we expect to have real systems that solve real problems with an associated lifecycle and so forth.

5. Effectiveness Measurement. Metrics for measuring quality and fidelity, for measuring success, for measuring return on investment, and so forth. We often consider these ancillary issues, when often they are the heart of the value. Without measurement there can be no management or understanding.

I Cheated a Bit…

While I said that I would like to start with 10 challenges, five technical and five pragmatic, I find myself compelled to add one more challenge. I shall conclude with this one because I think it is extremely important, and I believe it belongs under *both* categories:

Grand Challenge Problem 0: Public Benchmark Data Sets.

As a field we have failed to define a common data collection (corpus) and set benchmark measures relative to it. As a result, it is very difficult to judge both research and systems advances. The establishment of such a collection is not an easy task, but is also not impossible. We need a solid mix of both synthetic (but realistic) and real data sets in this collection. As a community, we must rise to meet this grand challenge and soon.

I would like to conclude this "longish" editorial. It is my sincere hope that through debate, feedback, and constructive criticism, you will all help me refine this initial proposal to a healthy and acceptable list of grand challenge problems. Of course, even if we later prove that we were way off in our thinking of what are really grand challenges and what really matters, then that in itself is an advance.

All in all, I remain extremely bullish on our field, on its prospects, and on the ability of our growing community to contribute significantly to the important research and application problems facing us. Much is hanging in the balance, and I can't wait to see the next set of advances in the field. My hope is to see much of the debate and discussion around this topic appear here in this newsletter. Of course, I also hope that *Explorations* will chronicle these challenges and the solutions that are advanced to address them. We certainly live in exciting times for our field. I look forward to great achievements by this community.

Preview of Current Issue

We are proud to present this special issue on Microarray analysis compiled by our special issue editors Gregory Piatetsky-Shapiro and Pablo Tamayo. They have gone beyond the normal duties of special issue editors in preparing this extensive collection of articles. We would like to thank Insightful Corporation for partly sponsoring the expenses incurred in printing this extra large issue.

An exciting feature of our December issues is the 2003 KDDCup articles where winning entries present a technical description of their approach. This is particularly valuable because this newsletter is the only place for finding such information. We would like to thank the cup organizers, Paul Ginsparg, Johannes Gehrke and Jon Kleinberg for their effort in compiling this part of the special issue.