Summary from the KDD-03 Panel -- Data Mining: The Next 10 Years

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

Gregory Piatetsky-Shapiro
KDnuggets
www.Kdnuggets.com
gregory at kdnuggets.com

Ramasamy Uthurusamy
General Motors
Detroit, MI, USA
www.gm.com
samy at gm.com

.

1. Introduction

This article is a summary of the panel held at the 9th International Conference on Data Mining and Knowledge Discovery: KDD-03 on August 27, 2003 in Washington, D.C. The panel participants included the following:

- Usama Fayyad of DMX Group (Panel Organizer)
- Rakesh Agrawal of IBM Almaden Research Center
- Daryl Pregibon of AT&T Shannon Laboratories
- Gregory Piatetsky-Shapiro of KDnuggets
- Raghu Ramakrishnan or the University of Wisconsin – Madison
- Ramasamy Uthurusamy of General Motors

The goal of the panel was to gather representatives from academia and industry and to ponder where the field stands after nearly a decade and a half of KDD meetings. We all have seen a significant growth in demand for data mining technology driven by a glut in data. We have observed data mining growing as a healthy research community. However, we still struggle on two important fronts: the scientific and the commercial. On the scientific front, Data Mining still needs to reach a stronger level of attracting steady contributions from the related fields. On the commercial fronts, the huge opportunity has not yet been met with adequate tools and solutions. This panel was an attempt to address the possible future directions for Data Mining and KDD.

Questions that the panelists considered prior to this panel included: Will we continue a healthy evolution to being a scientific field of study with a healthy contributing community? Will we go more down the path of systems and engineering? What are the next challenge problems? What are the milestones that define healthy growth and significant advances? Is data mining destined to continue to be a visible area of focus and research, or will it evolve towards embedded technology studied as part of other systems? The presence of a significant set of research challenge problems against which measurable progress can be made is a crucial component for the growth of a scientific field. What will these challenge problems be for KDD and Data Mining over the next 10 years and beyond?

Below is a summary of views of the panelists. We follow it with a summary of some of the topics discussed during the panel.

2. Position Statements

Position Statement by U. Fayyad (DMX Group)

The field is still vibrant, and the level of interest in it seems to be increasing dramatically, but many serious challenges lie ahead of us. My biggest fear about the field is the slow pace at which we are seriously engaging researchers from different fields. This is perhaps normal in the beginning, but I worry whether we have what it takes to change it in the future.

I think that the biggest stumbling block from the scientific perspective is the lack of a fundamental theory or a clear and well-understood statement of problems and challenges. This situation is in contrast to the very intense level of interest we are getting from practitioners and people interested in applications. These folks come from a large variety of folks. But without the scientific foundation and the continued academic progress, the field cannot grow and solidify.

Some of the other worries I have, as a data miner, revolve around the following observations:

- In a typical data mining session, I spend most of my time extracting and manipulating data, not really doing data mining and exploration
- The trail of “droppings” I leave behind in any given data mining session is enormous, and it seems every time it is replicated and repeated, almost from scratch, again.
- There is little theory, or even engineering practice, to what we do: It is all a "black art" today. We need a theory for what we do, and how we do it. The theory can drive engineering recipes so we can teach the craft to others effectively
- Without the participations of the relevant communities, we are doomed to a fate of badly re-inventing wheels, not a very bright future. How do we get people from other fields to engage?
- The core problems, we are too fuzzy about them. We need to start stating them crisply and we need to put the effort into showing other fields how our problems are fundamental to them and of great scientific interest to them
- Industry/environment: very encouraging, with some signs for worry.

Let me expand on this last point regarding the environment. The growth in storage and advances in hardware are what is driving demand for data mining, and unfortunately not our great success as a field. Stored data seems to double every 9 months, growing twice as fast as the infamous Moore's Law, and this makes demand for data mining and reduction tools increase exponentially, However, areas of particular concern include:

- The fact that no strong emerging standards exist for the field that makes adoption and exchange really difficult
- Privacy concerns are becoming a serious threat to our field

On the positive side, I see many applications and I see many companies recognizing the pain associated with doing data mining and trying to make the process easier. The market is growing because data sets are growing. If we effectively develop the science of data mining, tools, companies, products, and services will emerge to do it right because the market is there.

A list of grand challenges for data mining is important and should be developed. Primarily, we need these challenges on the scientific level. There are plenty of applications and challenges on the applications side. A list of these grand challenge problems, at least in my opinion, is a high priority for the field.

One such initial proposal for a set of Grand Challenges in the field is provided in the Editorial to this issue of *SIGKDD Explorations.*

Position Statement by G. Piatetsky-Shapiro (KD Nuggets):

Data Mining – the Past 10 Years

The most significant technical development of the last 10 years was clearly the World Wide Web. It ushered a new era in communication and processing information, dramatically increased the rate of information growth, and created many new fields, including web mining, with its own subfields of web content mining, web usage mining, search engines, bots and intelligent agents.

After the excitement and growth of 1990-s, followed by dot-com hype and bust period of 1999-2001, the data mining field is reaching a plateau of maturity. We can see it in attendance in conferences and state of health of data mining related companies. One indicator, the number of subscribers to KDnuggets (the leading newsletter on Data Mining), has been growing fast in 1996-2001, but remained essentially flat in 2002-2003. This plateau, however, is likely to produce new growth in several areas.

The few e-commerce companies that have survived the dot-com crash and are doing well, such as Amazon, eBay, and Google, all have active data and web mining programs and are hiring data mining experts.

The Grand Challenges Today

Here are some of the technical grand challenges for data mining today:

1) "Drop-in" classification. Drop in data set(s) and system produces a good classifier (including data preparation)

level 1: for one domain

level 2: for 3 different domains, e.g. banking, genomics, fraud detection

2) "What's New and Different" report. What is new in this data compared to past data, past knowledge, past expectations? Good progress on this topic was made recently with WSARE system (Wong et al, 2003)

3) A General Theory of "Interestingness". What makes this rule, pattern, finding more interesting than another? Many theories have been proposed here, including objective statistical measures of validity and surprise, and subjective measures based on user preferences and domain knowledge. However, no general theory has emerged yet. Perhaps, one part of the problem is that what is interesting is intimately tied to domain knowledge and domain, and instead of one general theory we may find that dozens of little theories would be adequate.

Social Issues

From a social point of view, the hottest technology and society topic in 2003 year was the controversy over the US government efforts in using data mining for anti-terror activities. While data mining has been just one component of those activities, the press frequently used "data mining" when used in press has been mostly associated with the unfortunately named Total (later Terrorist) Information Awareness (TIA) program,. See (headline from PC World: PC World, July 18, 2003: Senate Kills Data Mining Program www.pcworld.com/news/article/0,aid,111626,00.asp; Washington Times, July 18, 2003: Senate votes no to data mining funds http://washingtontimes.com/upi-breaking/20030718-070600-8337r.htm)

Even though Congress has closed TIA, we have not seen the end of this controversy, because there is an inherent tension between intelligence agencies that see an effective tool in data mining-based application and civil liberties advocates who see only an invasion of privacy.

Too often we have seen a very naive criticism of such programs, e.g. if there are 5% errors in data, and 300 million Americans, then there would be 15 million (=5%*300 million) of false positives. Such criticism ignores the power of repeated observations and link analysis in reducing false positives to acceptable level.

People are willing to accept false positives identifications and some invasion of privacy if they believe that this helps reduce the risk of terrorism. Consider that millions of air travelers in 2003 have removed their shoes and all of them were false positive. Yet few complained about that obvious erosion of privacy (e.g. a hole in a sock was revealed), because most accept shoe inspections as a way to prevent another shoe-bomber.

In addition, there are a number of technical solutions for preserving privacy while data mining, such as randomizing results, removing identifying information, etc (See Technological Solutions for Protecting Privacy, Roberto J. Bayardo, Ramakrishnan Srikant, IEEE Computer, Sep 2003)

I believe that it is important to take a balanced view of this issue and consider potential effectiveness of analytic programs and weigh potential benefits versus erosion in privacy.

The Next 10 Years

Hot applications for data mining next 10 years:

- **Text and web-content mining**: Web mining will reach new level with XML and Semantic web applications. Web is already a vast repository of useful knowledge, but with advances in intelligent agents, widespread XML adoption, and web services, web mining will reach the semantic level.

- **Relational mining and link analysis**. This will have applications in many fields, including biology, business, library and information science, marketing, security and perhaps will create completely new areas.
- **E-commerce.** We already see that the successful ecommerce companies like Yahoo, Amazon, and eBay are investing in data mining in order to extract value from their data. We expect these efforts to continue and expand also into intelligent bots.
- **Bioinformatics and in silico drug design**. this includes analysis of genomic data, DNA microarrays, etc. In silico drug design is still a very promising area. Successful results have already been reported, e.g. highly accurate microarray-based diagnostics.
- **Multi-media data mining.** We already can find images and video based on a keyword. Combined with image recognition, we can envision many interesting tasks based on actual understanding of images, video and audio.

Finally, I expect the social issues of data mining, threat detection and privacy will remain a significant concern for data miners and for the society at large,

Position Statement by Daryl Pregibon (AT&T Research)

The challenges and opportunities facing the data mining community are greater than ever, yet it is very difficult to claim that the effort has yielded much success. Even worse, the pejorative connotation that data mining had in statistics (it ain't science) is now almost universally held in society (tools to invade my privacy).

In order for the field to survive the short term, and build for the long term, another threat needs to be addressed, namely the fickleness of the research community. There are, to varying degrees of activity, several emerging fields that are inherently data mining activities, e.g., discovery informatics and network forensics. These have the potential to peel off leading researchers from the data mining field, and more importantly, attract the best new researchers. There is the very real possibility that data mining will go the route of expert systems some twenty years ago as the community jumped onto the neural networks bandwagon (gravy train).

KDD has a history of being very inclusive, and long-term viability will depend on this model even more in the coming years. The field needs to be perceived by the research community as vibrant, exciting, leading edge, and fun. But history contains many sad stories of what happens to leaders that are no longer perceived to lead. KDD should take note lest we become a mere footnote in the march toward fully realizing the goals of knowledge discovery and data mining.

Position Statement by Raghu Ramakrishnan (U. of Wisconsin – Madison)

Data mining has several challenges and opportunities ahead of it, and I'll bring up some that I think are important, but with no claim that these are the most important issues. My crystal ball is a little cloudy...

I see three distinct kinds of challenges: Technical, social, and economic. The technical challenges consist of identifying interesting research thrusts that keep the field vigorous and moving in high-payoff directions. The social challenges consist of how to rise beyond the perception that we primarily enable junk mail and other, more serious, intrusions of privacy. The economic challenges consist of making the field pay for itself, and achieving widespread recognition of this fact. We are at a cross roads where the initial enthusiasm has steadied, and these challenges are squarely in front of us.

The technical challenges that I plan to highlight are:

(1) Finding ways to address the real bottleneck in data mining, which is the human cycles spent in analyzing data. Fast algorithms are sexy, but real advances will come from techniques that will lead to more efficient management of the process of data mining, and that reduces the cycle time in arriving at useful insights.

(2) Data mining is often perceived as a bag of tricks. We need to at least provide a vision of how these tricks fit into a coherent tool-kit. In the process, we need to understand the scope of the field---machine learning, statistics, databases, information retrieval, multimedia analysis---and how to create synergy between the different communities that intersect here, rather than isolated pockets that occasionally share a conference venue.

The social and economic challenges are closely related---if we can enable significant use of data mining techniques that are seen as positive applications of the technology, and that generate significant revenue, we're in good shape. If not, we have a problem. Can KDD researchers directly influence this aspect of the field? We need to consider this, and take concrete steps in our research, our outreach, and our peer review process. While these are "soft" considerations that we as engineers and scientists often feel uncomfortable quantifying (which means "evaluating", for most of us!), these are nonetheless difficult challenges that will have a much bigger impact on the field than almost any technical difficulties that we face.

Position Statement by R. Uthurusamy (General Motors)

A casual surfing of the web yielded two extreme views of the KDD field. One vendor declared (1997) "KDD is obsolete." In the other (1998) Arno Penzias claimed that "If you are not doing KDD, you are out of business." I suppose the truth of the state of KDD lies somewhere in between. Here is my assessment of "KDD: Past, present, and future."

KDD started out to be an application-driven multi-disciplinary field, and still continues to be so, and I expect it will also be so in the future. It may not ever become a pure scientific area in its own right like Statistics or Database (DB) or Machine Learning (ML). This might not have been the original or even the current intent. The slow progress thus far and other distractions caused by the web and vendors to its efforts are moving it away from its possible and potential scientific quest. While fifteen years is too short a time to have a great scientific impact, KDD did make progress and accomplished the following:

KDD created a sharp discontinuity in economic affairs and opened up opportunities for businesses to do things with data that had not been done before, and allowed things to be done in new ways.

Three fundamental issues distinguish it from other fields. First, it did bring awareness to the scientific community, the need to deal with and make sense of ever growing data that the industry and business communities are facing. It is not just the very large size of the database, but also the very high dimensionality of the data that demanded new research efforts. Secondly, it also emphasized the need for a discovery component, the fundamental part of any KDD application and solution. Thirdly, it attempted to steer practitioners towards a process and system-centric view of KDD for it to be effective in solving real world problems.

- It had been a catalyst in educating the professionals on the need for data cleaning through the discovery that only ten percent of collected data is ever used and forty percent of all collected data have errors in them.
- It had accelerated the need for and the development of data warehouses and the requirement of a team of consultants who understand the KDD process and for a team of multi-disciplinary experts for a systems-based approach to solutions.
- This awareness actually was a catalyst in getting the ideas and techniques of KDD used not by people and businesses that had and now have to deal with massive data but by small and medium industries and retailers who have relatively small and medium databases, but who saw an opportunity to make better sense of and effective use of this data.
- It has also spawned many new industries and new research areas apart from the original KDD focus. Some examples of these are CRM, web mining, text mining, collaborative filtering, database marketing, etc.
- KDD is not one-to-one marketing or mass customization or even database marketing termed as relationship marketing. Original KDD efforts and even KDD papers and articles paid very little attention to these areas, which turned out to be a major mover of this field and sometimes even mistaken for it.
- KDD has seen only incremental advances in tools and techniques. A fragmented focus on bits and pieces of the field had resulted in some disappointing progress in the following, not just because KDD issues were hard, but because of less attention being paid to these. Some of these have been left to other fields to pursue.
- Some notable difficulties are in getting a handle on what are K, D, KD, and KDD in KDD. What is more disturbing is that some KDD practitioners don't seem to care much on this when compared to the original zeal on these topics. If the field does not seem to care for "the reason for its being" and "specify explicitly what it is about," there is the danger of it being pulled into many different avenues.
- The key issues of "interestingness," "privacy," "standards," "data cleaning," and more importantly the "process and system centric approaches," have been recognized but have seen very little progress towards addressing them. These were also the original concerns of the field but the current trend indicates that they will not be addressed well for a long while.
- Even obvious and immediate needs like visualization, massively parallel computation, and the core need of integration have not received their much-deserved attention, even though they have seen some advances. The original process centric view of KDD espoused the three "I"s (Integrated, Iterative, and Interactive) as basic for KDD. These are central to the ideas of "Computer Assisted Human Discovery" and "Human Assisted Computer Discovery." There has been very little work on these in recent years.
- The issue of evaluating the discovered knowledge and making it actionable was also part of the original concerns. More work is warranted here. More so because, to do it right required collection of more data and analysis where KDD itself becomes the driving force. This difficult area demands more research.
- Even in successful areas like database marketing, one-to-one marketing, customer profiling, etc., the focus has been on reducing cost and improving the profit of the firm doing these. While this is a start, further development of tools and techniques are needed to focus on the customer, to whom these are targeted, who will benefit by all these mining of data and focused marketing.
- A key disappointing lack of progress is in getting the Database, ML, and Statistics communities to work closely and cooperate to focus on solving KDD issues. There have been efforts to foster their interaction, like connecting KDD conferences with Stat and ML conferences, joint talks and sessions, and summary sessions. These have not yielded the expected collaborative efforts. Even the gathering of KDD luminaries in 1997 in Seattle along with luminaries from the Statistics, Database, and Visualization communities for a KDD summit has not produced any measurable and notable outcome. I did request them to focus on system and common issues but to no avail. If much closer cooperation and interaction do not result soon, the KDD field will remain fragmented with occasional interdisciplinary contributions. The interaction issue will become much more difficult and complex due to the current KDD trend demanding closer interaction from additional fields like Information Retrieval (IR), Natural Language Processing (NLP) and Understanding (NLU), and the web community. My earlier attempts to get these areas well represented have only been a start but this needs to be accelerated.
- Vendors have not yet provided truly integrated KDD tools and thus leaving customers with unfulfilled expectations. The education of the user community in the KDD process and what is and what is not possible becomes paramount.
- The major advances that benefited KDD in the past ten years were not directly due to KDD needs but were the

results of other pursuits and objectives. ID3/C4.5, MARS, and Association Rules are examples of contributions from other areas that have been instrumental and influential in KDD's growth. Neural Networks and Visualization areas had their share of contributions to KDD as well. But, KDD itself needs to contribute scientific advances that are in line with these in terms of impact and use.

As far as the future of KDD, especially its scientific progress, is concerned, it seems easy to predict what will happen in the next two or three years. These short-term trends will make predicting long-term contributions very difficult.

External influences like the phenomenal growth and use of the web and some vendor hype will in the short-term cause distractions to the field. They will cause the field to focus more efforts on CRM, Database Marketing, OLAP, etc., than on efforts that will cause fundamental scientific KDD advances. The basic research community needs to ignore these temporary distractions and focus on solving the right KDD problems. This brings up the issue of what will happen versus what should happen. The trends in KDD by necessity depend on the trends in its component science and technologies that also depend on progress in related areas.

The definite trend seems to be the split of KDD efforts into two parallel areas of progress. Both will demand certain scientific advances and both will meet some real world needs. I call these two GM-way and Microsoft-way. The Microsoft-way will move the field towards "data mining for the masses." The GM-way will move it towards "KDD for business and industry," where massive data is easy to collect but difficult to exploit and use. The two have parallels in mainframe computing and desktop computing.

Looking back ten years, I see there were really no true predictions of things to come, but wish lists. In the interest of making the point, I might take the extreme view and predict that none of the following important KDD issues will show any measurable and notable scientific progress!

- Process and System-Centric KDD
- Well integrated and closer cooperation among its constituent fields DB, Statistics, ML, IR, AI, NLP, . . .
- Assisting users on how to select data as well as appropriate and relevant methodologies that align with the properties of that data
- Interestingness
- Incorporating domain knowledge

Apart from the above Grand Challenge problems to be addressed by the KDD community, the issues facing the field from a social/utility perspective are:

- Educating the public and government on the *KDD Process*
- Privacy and Security
- Scalability
- Mining Data Streams
- Standards
- Minimizing the hype

I envision the convergence of Semantic Web, P2P, Internet2, Grid Computing, Web Services, and Ontological Engineering that will lead to interesting new avenues of KDD research. More importantly, the KDD community should pay attention to "KDD for Strategic Intent" that I believe to be a very promising high impact business application area. An example is how Harrah's achieved significant growth and gains over competing casinos by formulating strategies based on the results of mining customer data. The concept of utilizing KDD tools and techniques to formulate corporate strategies that provide competitive edge is at the core of the intent of KDD to start with.

KDD started as a field by integrating results from other fields like ML, DB and Statistics but it needs to make many true contributions of its own. KDD will survive because it solves some immediate business problems and provide measurable benefits and it will do better if the above basic issues are addressed. With today's tools and techniques, it is the case where KDD, while drowning in the data glut, is drowning in opportunities.

3. The Panel Discussion

The panel consisted of expansions on the various position statements listed above. This was followed by a series of question/statements from the audience. The highlights of the discussion section of the panel that are not captured in the above position statements are summarized below.

Rakesh Agrawal chose to present in the context of the KDD-2001 panel discussion on a similar topic. He presented a series of comparisons between the problems that the panelists presented then and whether, in his judgment, there has been significant progress over the past 2 year. He noted that there has been little progress made on the front of defining what data mining is and evolving a theory behind it. On the front of privacy, he concluded that some progress is being made on the privacy-preserving data mining, and that this remains an area ripe with research and can benefit greatly from randomized algorithms and existing research in cryptography. One of his primary themes is that it is a mandate on the field to solve the problems of privacy and data mining. (Note: Rakesh Agrawal has recently been honored by Scientific American as one of top 50 leaders for his work on privacy and data mining).

On mining on the web beyond simple click-stream and simple html analysis, Agrawal noted that there has been some good recent progress as evidenced by papers, but it is too early to judge whether this progress is sustainable and significant.

On the front of mining for actionable patterns and including domain knowledge, some serious progress has been made, especially in classification, but we still have a long way to go. A similar prognosis holds for several other interesting technical problems presented in KDD-01. On the front of what problems not to work on, the situation remains the same: it is too early to judge, as we still know too little.

Agrawal concluded by listing his grand challenge as an augmentation to the prior discussion on mining several data sources simultaneously (to which he added new data sources to the 2001 list). His Grand Challenge is summarized by: *Find What's there, What has changed, Across sovereign data repositories.*

R. Uthurusamy gave a colorful presentation drawing on a multitude of analogies. He drew on several industry examples including MCI, Harrah's, GM, and some others. He concluded with material from industry analysts covering the various “hype” curves for new technologies and argued that KDD and Data Mining are still in the very early stages despite what we think has been a lot of unjustified hype.

R. Ramakrishnan discussed the challenges for KDD along three major fronts: technical, social, and commercial. He highlighted in particular the challenge of “*Faster KDD*” versus “*Faster Algorithms*”, focusing on the fact that the human factor in mining needs much attention. The other major challenge is in what he called “*Data Minding*” *systems:* systems that monitor data streams and can maintain themselves and their models of the world, and automatically manage and retire/purge data and models as appropriate. He concluded on a warning note that we as a field have a chance to either disintegrate into separate communities, or come together and contribute to the problems as one field.

4. Concluding Remarks

While the job of predicting the future is risky at best and likely a losing proposition, we can learn from discussing and analyzing the challenges of today. At the highest level, the field is facing healthy and interesting challenges. We do not believe any of them are insurmountable. Besides, the opportunity is so large, the need so great, and the advances so far so small in comparison, than any advance going forward will stand a great chance of being significant. The community is healthy and stands close to a crossroads. The big “social” challenges lie in overcoming issues such as privacy concerns and “bad press” and also in the KDD community figuring out a way to communicate its problems crisply and clearly to other fields. Without active contributions from the other related fields, we stand to miss out on important advances and achievements, and we as a community will be doomed to re-discover and re-invent solutions to problems that have been tackled before. On the technical front, the wealth of problems and challenges are sure to keep us busy as researchers and applied practitioners.

One interesting thought is to work on what appears to be a very effective approach to address both technical and social challenges is to formulate a set of grand challenge problems. An initial proposal of what these Grand Challenges might be is provided by Fayyad in the Editorial to this issue of *SIGKDD Explorations*. These challenge problems can serve as a beacon for researchers in the KDD community as well as a bridge for contributors in other fields who can spot and understand these problems. Some of these challenge problems were discussed in the panel. We will follow up with a more crisp statement of the challenge problems along with an endorsement from members of the KDD community that this list of problems is worthwhile and significant. We should all be thankful that we work in an area that is rich with open problems.