KDnuggets Interview with Usama Fayyad

Gregory Piatetsky-Shapiro
KDnuggets

gps@KDnuggets.com

ABSTRACT

The KDnuggets newsletter has a new section of interviews with leaders in the field. This article presents the interview with Usama Fayyad, President and CEO of digiMine.

1. INTRODUCTION

In May 2001, KDnuggets inaugurated a new section devoted to interviews with leaders of the field. I am pleased that the first one is with Usama Fayyad -- a researcher, an entrepreneur, an industry leader, and a friend.

I first met Usama in 1989, when I tried to hire him for a summer job at my project on Knowledge Discovery in Databases at GTE Labs. Fortunately for the field, he declined and went to JPL where he made his first major "out-of-this-world" discoveries. Usama has remained a friend and a colleague since then.

In this interview, Usama answers questions ranging from his work at digiMine, to his opinion of Bill Gates (whom he met at Microsoft), and to predictions on which companies will succeed and fail. Usama's brief bio follows the interview.

2. Interview

GPS: *What made you interested in Machine Learning and Data Mining ?*

Usama Fayyad: I've always had a great love for mathematics, especially probability theory and reasoning under uncertainty. Later, I became intrigued by computers and the notion of digital data and digital worlds. Data mining, pattern recognition, and learning algorithms are areas in computer science where mathematics can use data to answer unanswerable questions and solve seemingly impossible problems.

Data mining is the ultimate detective science; evidence is plentiful but hidden in mountains of data. It's an exciting field. I don't know how anyone could resist it.

GPS: *What is your typical daily schedule?*

Usama Fayyad: Every day is different and exciting, and it really depends if I'm traveling or not. If I am in the office, I get in very early and use the time to catch up on e-mail. I process a good 500 or more e-mails each morning. I am on almost every group alias in the company. It is a nice way to tap into the pulse of digiMine passively and getting a feel for what is going on in sales, in R&D, in marketing, in finance, and in operations. I know when something is going wrong and I am never surprised. My days in the office are usually packed with meetings except for certain slots that my assistant, Cathy, reserves as time to think and catch up. I constantly meet with various digiMine groups and managers, as well as visiting customers and vendors. Throughout the day I field numerous phone calls from customers, partners, employees and investors. Our investors are extremely active and engaged in digiMine's business, especially when it comes to sales and marketing.

At digiMine we serve meals for lunch and dinner, and this gives me a chance to catch-up with employees in the cafeteria, grab some food and work through lunch or dinner. I am very big on continual and hierarchy-free communications. So three days a week involve direct interaction with various employees; our weekly all-hands meetings attended by over 100 people. We have a weekly "Drinks and Stories" session on Friday evenings where we exchange fun stories and I answer candid and tough questions. And we have a weekly lunch session with a group of 7-9 employees intended to foster more personal interaction and maintain a flat organization. We believe in open management and pretty much share all numbers and facts about the business with all employees. We think this builds trust and a sense of control and ownership for the employees.

I travel a great deal. On travel days, the pace is even more hectic. I love red-eye flights to the east coast, or super-early flights to the west coast. It makes efficient use of time. I hit the ground running as my travel days are packed with back-to-back meetings. Most are with new and prospective customers. I love meeting customers. Some meetings are with VC's, banks, investors, or analysts where I cover the strategic aspects of the business. I enjoy the fast pace and the challenge of switching mind-set from meeting to meeting.

Every day is full of new and different challenges. Surprises and change are inherent to managing an emerging business. Major product issues arise, competitive threats emerge, or customers present unforeseen issues. It's a roller coaster and it's not for the faint of heart. But the thrill of solving the problem, winning the account, or outmaneuvering your competition makes it all worthwhile. Then you feel on top of the world. I used to think nothing could be more exciting than down-hill skiing a tough slope, but I think my life at digiMine is much more exciting.

GPS: *At JPL/NASA, you developed SKICAT project that made out-of-this-world discoveries by analyzing star data. What was the most innovative part of your SKICAT work?*

Usama Fayyad: I believe the most innovative part was that we used data sets that were collected for a different purpose to solve a fundamental problem in the field. The sky survey is done in a certain image resolution and is intended to cover the entire visible sky with a few thousand plates. However, very high resolution images of a tiny part of the sky covered by each photographic plate are taken by a separate telescope to help calibrate images and pixel intensities across multiple parts of the sky and multiple times of the year. The big insight was realizing that an astronomer can readily recognize a sky object in the high-resolution image but see nothing but a few unintelligible pixels in the regular survey plate image. The challenge then was to find out if the few low-

resolution pixels contained latent information sufficient to allow a machine to match the human recognition in the corresponding high-resolution image. The success of the mining algorithm at extracting this accurate classification capabilities from measurements done over the few low-resolution pixels allowed us to be able to outperform highly trained human astronomers on images where no high resolution views were available. This looked like magic to astronomers. We wound up solving problems that astronomers struggled with for decades. All thanks to algorithms that could take advantage of a large number of measured variables per sky object. It turned that the classification was inherently high-dimensional, and hence a human-based analysis approach, using traditional statistical analysis techniques, was not going to yield the required accuracies.

That is an excellent example of a classic data mining application: lots of data, a difficult analysis task, and users who are very motivated to work with you to solve the problem. This kind of settings lets data mining algorithms shine on aspects they excel at: the search for useful dimensions out of a large set of possibilities.

GPS: *After JPL, you joined Microsoft Research, where you led a data mining group that developed a number of data mining components for Microsoft Products. What are some of the difficulties and successes of translating research into product development?*

Usama Fayyad: I started out at Microsoft Research and grew the Data Mining & Exploration group. The charter was to continue to do basic science. At one point, we also decided to form a parallel product group to ship the data mining components and a new API as part of SQL Server 2000. This actually was a crucial and important transition: to form the product group and focus it on development. It required on my part the ability to abandon a lot of the freedoms that were the perks of the research life and take on the responsibilities associated with shipping a product. It meant living with the product team and relocating to their building, etc.

The major difficulties in getting the fruits of research into a real industrial product all revolve around how to change the perspective and the definition of the technology so it becomes something a development group can live with and work with. For a company like Microsoft, it was key to understand that for a platforms company, pushing data mining algorithms and techniques was the WRONG thing to advocate. I had to come up with a new view of data mining as a reasonable and natural extension to the platform. Hence, we could not build tools that are intended for specialists and Ph.D.'s. We had to figure out how to package the technology so that a developer building on the database platform could use the technology in a natural manner. This also meant integration of the requirements in the platform and letting the system manage the scary details.

For example, in data mining you create a model, you train a model, you update a model, you apply a model, and you delete a model. The magic contribution of working with a product group was to get them to figure out the normal analogies. CREATE statement in SQL does exactly what you need: define the structure of a model, and then make it a first-class citizen of the DBMS, just like a table. Training was akin to INSERT INTO where you insert data into a table. Applying the model, say for prediction, was akin to the JOIN operator. This was a whole new way of thinking about data mining operations. But in the world of SQL these were familiar and understood notions. That is one of the examples of notions that you never consider when you are in the advanced research world. But they are crucial if the technology is to be adopted by product groups.

GPS: *What is your impression of Bill Gates?*

Usama Fayyad: Bill is a very smart individual who has an amazing capability to focus intently on different topics. It amazed me that when we did a briefing on data mining, he would focus on the details and ask very specific and relevant questions. The reason this is amazing is that he does these briefings with many groups at Microsoft. The capability to move the focus of attention from one area to a totally different one, without being distracted, and with strong analysis, is a great talent.

The other great respect I have for Bill is that he actually takes the time to sit through product and technology reviews. Detailed ones. All this while he was still running the most-valued company on the planet. He also has a great ability to pick very smart and capable executives around him. Steve Ballmer is a great business leader. Not just very smart, but an amazing motivational and inspiring speaker. Rick Rashid (Sr. VP of Research) is also a super-smart and accomplished academician who is in tune with the business issues and strategies. Microsoft had an impressive bench of powerful and smart executives. Many of them wound up early investors in digiMine after they left: Pete Higgins, Sam Jadallah, and many others :-)

What I admire most about Bill Gates is that he managed to build a clean, agile, and effective organization. His recent relinquishing of the President and CEO duties are admirable as they came at the right time and showed that he can make the tough business decisions. Ballmer was a great choice for succession.

GPS: *How did you decide to start digiMine?*

Usama Fayyad: Late 1999 and early 2000 was an environment where venture capital was very accessible. I would get calls from many VCs saying that if someone like me were to start a new company, funding would be instantaneous. Psychologically, it is always scary to give up a dream-job at Microsoft Research, to go out and start something from new from scratch. The scariest part was sacrificing family life and giving up all the security and the huge amounts still locked up in unvested stock options.

Two major factors had an overwhelming impact on me. One was personal and one was professional. On the personal front, I had achieved success in an academic research institution - Caltech's Jet Propulsion Lab - and had won NASA awards and medals. I also achieved success in an industry research lab: Microsoft Research, having built a basic research group and then a product group. I had shipped technology in 3 major products ay Microsoft. I felt I also experienced the world of technical publications, having chaired several conferences, including KDD, and being Editor-in-Chief of the primary technical scientific journal in Data Mining. I had also co-edited the KDD book [and many articles with you Gregory]. What I was missing was the element of risk. The feeling that everything is on the line and that what I do next could make the difference between success or utter and disastrous failure. The "security" of my position was dulling my senses, and I needed a new sense of risk and a true thrill of the hunt. Doing a high-tech start-up was definitely the next thing. And indeed, the capital materialized instantly and digiMine grew very fast from three cofounders to over 100 people in less than a year.

On the professional side, I was watching companies really struggle with data warehousing and data mining. Companies could not effectively run and maintain data warehouses, warehouses were overly complicated and expensive, and the business users were not getting any value from the data. Consequently, running a data mining application was essentially out of the question for an overwhelming majority of all businesses. To run a data mining application, you had to embark on a multi-month data quest. We figured the only way to get data mining to work, was to simultaneously solve the data warehousing problem and integrate business solutions within it. Key to this vision was running the data warehouse as a fully-hosted, fully-managed service for companies. The technology is so complex that a hosted model was the only way to make it work, to make it economical, and to make sure business users get the benefit.

GPS: *What is your vision for digiMine? How it would become profitable?*

Usama Fayyad: The vision of digiMine is to make data warehousing, mining and business intelligence usable by business people and more affordable by businesses and organizations. Currently, many companies are prevented from tapping into the valuable intelligence their data holds because managing and analyzing the data is too complex and expensive. We want to resolve that by taking on the complexities of data warehousing and mining in a hosted service, and presenting the analytics to business people in a format they can understand and use. We also package and build business applications that utilize data mining and the data from the warehouse: real-time cross-sell, up-sell, prediction, and segmentation technologies. There is a tremendous market for this type of solution, across many industries and all sizes of companies. Our plan at digiMine has always been to be profitable by the end of 2002 and we are still on course to reach that goal.

The solutions we offer can cost 10 to 100 times more if companies attempted to build them in-house instead of using digiMine's offering. They also would require hiring very rare talents: database developers who understand data warehousing and data mining. Large data operations teams to manage the ever-evolving data warehouse are hard to build and keep. The trick is to balance growth and expansion into multiple market segments, with the need to meet profitability goals.

Of course, we all know that at the end of the day, technology and a great business model are not enough. Marketing and Sales play a huge part. Getting market mind-share and awareness is a crucial challenge. We've had an amazing and outstanding start, but we need to keep executing to meet our goals.

GPS: *What is a recent book that you read and liked?*

Usama Fayyad: The Monk and the Riddle, by Randy Komisar. It's an insider view of the world of Silicon Valley VCs.

GPS: *Excluding your own work, what were the largest successes of data mining so far? What were the biggest failures?*

Usama Fayyad: I think some of the early applications that IBM did between 1996 and 1999 have been very important to establish credibility for the field in business settings. Some of the application companies, such as SGI, NCR, and many others drove serious long-term advances in the field by productizing the technology. The most widespread impacts on the field will come from efforts to standardize the way data mining is used, invoked, persisted and shared. This includes some of the IBM work, some of the Microsoft work (OLE DB for Data Mining), some of the Oracle attempts, and groups like DMG and the PMML standard for predictive models.

The other big development is the standard statistical package providers beginning to offer data mining tools. This includes S+, SPSS, SAS, and others. They still are complicated tools, but it is an important signal to statisticians and practitioners that data mining is something that needs attention and is getting attention.

The field now has evolved beyond the need for basic technology, like trees, clustering, SVM's, and so forth. The strong need now is in figuring out how to do applications that work, that scale, and that are easy for people to understand as business solutions. Work in context of business solutions is likely to bring through the next breakthroughs. The science must continue as well, of course, to facilitate the business solutions. Science and business need to work hand in hand to push the field forward.

GPS: *Currently, many dot-coms and data mining companies are having a hard time. Which companies will recover? What is your prediction for NASDAQ a year from now?*

Usama Fayyd: As every data miner knows, the prediction business is tricky. Only reasonable predictions come with attached uncertainty interval. If I were to attempt to predict the NASDAQ, My expected variance is fairly large. But since you asked, I'll take a quick attempt: I believe that the real historical growth rate should be between 10% and 12% per year. Technology companies are still commanding higher PE ratios than stable companies, but that is justified because technology will continue to grow. Given the flat regime we're in now, I expect well be about 10% higher next year, so my range on NASDAQ is 2100 to 2300 a year from now. The biggest factor that will impact it is what happens to oil prices. If they come down, technology will come back up big time. If they rise from here, we could be in trouble.

As for dot coms and data mining companies; the two categories of companies are very different from each other. Data mining companies are technology infrastructure companies or business solution companies. Dot coms are content, commerce, or services portals. Dot coms received irrational acceleration in valuation. Now the pendulum is swinging the other way. Will the Internet continue to be crucial? Absolutely. It is too fundamental a change to disappear. Businesses and consumers are more connected to each other today than any time in history. The Internet infrastructure will fundamentally change the world. It has not even begun to yet. But change never happens overnight.

Data mining companies never really went through a big bump in valuations. I think we saw a flurry of activity with over 300 vendors in the field. These vendors are selling the wrong tools to the wrong audience (sophisticated DM tools to business end-users). They need to transition to selling solutions. They are learning. But many will disappear in the process.

However, for anyone in the data mining business, there are some very encouraging powerful forces in technology today that indicate that KDD and data mining are going to continue to grow in importance. The new "natural laws" in our digital data world paint a very bright picture for data miners. We are all familiar with Moore's law that says processing capacity doubles every 18 months. Few people are aware of a more aggressive cousin to that law.

Data storage capacity doubles every 9 months. This law has been in operation for over 10 years now. The results of the two laws are all around us, people have way more data than they know what to do with. The "natural laws" lead to a prediction that I am willing to bet all my fortunes on: the gap between how much data can be generate, and our ability to process it will continue to grow dramatically. This means that the need for technologies to help reduce, understand, mine, and exploit this data will grow in importance. Taking a simple processing approach to dealing with this data will not help. We need the next generation tools and regimes. This is a HUGE opportunity for data mining. It is up to us to make sure we respond to the opportunity by delivering the tools of the future. This is nothing less than building the bulldozers, the cranes, and all the power tools of the digital data universe. There is a lot of useful structure in data out there. Now it is time for data mining tools to help us discover it and exploit. The future is indeed very bright. However, no one knows what kinds of companies are likely to figure out the right formula to effectively benefit from it. I promise to continue trying to figure it out...

Brief Bio:

Usama Fayyad is President & CEO of digiMine, Inc. He received his Ph.D. from The University of Michigan, Ann Arbor in 1991. He is an Editor-in-Chief of Data Mining and Knowledge Discovery, the primary technical journal in the field, and has chaired several past KDD conferences.

Prior to digiMine, he was at Microsoft where he founded and led Microsoft Research's Data Mining & Exploration (DMX) Group. His work with Microsoft product groups included the development of data mining prediction components that ship with Microsoft Site Server (Commerce Server 3.0 and 4.0) and developing scalable algorithms for mining large databases and architecting their fit with server products such as Microsoft SQL Server and OLAP Services. In addition to managing the DMX research group, he managed the core part of the development team that is building providers in the SQL Server product group. He was also a driving force behind establishing a new industry standard in data mining based on Microsoft's OLE DB API.

Prior to joining Microsoft, Usama was at the Jet Propulsion Laboratory (JPL), California Institute of Technology (1989-1995) where he founded and headed the Machine Learning Systems Group and developed data mining systems for the analysis of large scientific databases. For this work he received the most distinguished excellence awards from Caltech/JPL and a U.S. Government Medal from NASA. He remained affiliated with JPL as Distinguished Visiting Scientist after he moved to Microsoft.

Usama lives with his wife and their 3 children in Seattle area.

About the author:

Gregory Piatetsky-Shapiro, Ph.D. is the President of KDnuggets, which provides consulting and recruiting services in the area of data mining, web mining, knowledge discovery and business analytics. Gregory is the Editor of KDnuggets News and associated KDnuggets.com website.