

Interview with Usama Fayyad, Yahoo Chief Data Officer

Gregory Piatetsky-Shapiro

I am pleased to present an interview with Dr. Usama Fayyad, conducted in October 2005.

This interview was first published in KDnuggets News (www.kdnuggets.com/news) ([05:n20](#), [05:n21](#), and [05:n22](#)).

Dr. Usama Fayyad is probably familiar to most data miners and KDnuggets readers. He has many outstanding accomplishments, including publishing many significant research papers and several books, co-founding KDD Conferences, ACM SIGKDD data mining society, and Data Mining and Knowledge Discovery journal, leading NASA and Microsoft research on Data Mining, and founding 2 companies (digiMine and DMX Group).

Recently he became Yahoo! Chief Data Officer (the first such title in the industry). Here is [Fayyad's bio](#) (docs.yahoo.com/docs/pr/executives/fayyad.html), which covers his many other achievements.

I first [interviewed Usama Fayyad for KDnuggets in 2001](#) (www.kdnuggets.com/news/2001/n11/).

Question 1. *Gregory Piatetsky-Shapiro: Congratulations on becoming the Chief Data Officer of Yahoo (first CDO in the industry) - How did you decide on this title and how difficult or easy was it to convince Yahoo ?*

Usama Fayyad: Thanks, Gregory. It feels good to be the first in what I have no doubt will be a growing trend in the industry; and it feels especially good because I think it bodes extremely well for our field of work: Data Mining and Knowledge Discovery in Databases -- showing its particular relevance to business and the great achievements of this field over the past decade.

My title was constructed jointly by Yahoo!'s CTO, Zod Nazem, and myself. Chief Data Officer described exactly what the executive team and the company were looking for: someone to lead all strategic data activities and to represent Data as a strategic asset that DRIVES business and that helps lead the company in new directions. I totally believe that most large companies will start appreciating this dynamic and we will see many more companies with this exciting and very much needed position.

What is great about the story of my title is that I did not have to convince Yahoo! that it needed to be created, quite the opposite...

Yahoo! basically approached me saying that they recognized the strategic value of data and wanted Data to have a voice at the executive table. This was a surprise to me as I had spent the last few years building up a business, DMX Group, whose primary objective was to explain the value of data to many CEOs and most senior executives in Fortune 500 companies. The goal was to build a bridge between data and business and to explain that data is a strategic asset, not an IT service.

Hence, Yahoo's executive teams (and founders) were ahead in their thinking and understanding of the value of data; more than any other company I met with. As Yahoo became a client of DMX Group, the relationship grew closer and finally Yahoo! acquired parts of DMX Group that were relevant to Yahoo!'s business (which included myself and a number of other DMX Group employees). Part of the acquisition condition was the ability to spin-off the remaining business of DMX Group and guaranteed that none of the clients or existing projects of DMX Group would be negatively affected.

During the months that I did consulting with Yahoo! while at DMX Group, I also learned a huge lesson: that the Internet and interactive on-line applications are a much larger opportunity than I ever imagined. This felt

very much like the "next wave" of data mining, and frankly the next wave of the new generation of business and global commerce. This made it near impossible for me to resist the opportunity to work at Yahoo!, and indeed I marvel at how much I've learned in the past year about the Internet and the advertising businesses.

Q2. GPS: What are some of the biggest data mining challenges you face now at Yahoo?

Usama Fayyad: Yahoo!'s users, through their use of our network of products, generate over 10 terabytes of data per day. This is the equivalent of the entire text contents of the library of Congress. This is data that describes product usage, and does not include content, email, or images, etc.

The first and largest challenge is the ability to capture all of this data reliably, process it, reduce it, and use it to feed the many, many reports, applications, and data warehouses, data marts, dashboards, and scorecards across the company and its businesses. This is a game of reliability and scalability. You cannot fall behind, because you can never catch up if you do. Because this data stream is always growing (Yahoo now serves over 410 million unique users a month!) you cannot just plan for the existing data load, but always be building ahead of the game. There simply is not enough time in the day to play catch up or reprocess the data. Also, this data comes from thousands of servers that are around the world, with new servers being added and old ones replaced all the time...

A second challenge is defining metrics that are central to the business and understandable by the business units. This is a very tricky area. Yahoo! is in a wide range of businesses and verticals. Figuring out how to process the data and present the results in ways that are actionable by the businesses is not easy, especially in the Internet space where things change fast and on an ongoing basis. This also includes keeping up with new pages and new products being launched almost on a daily basis -- this is an environment that is very far from static! In addition, it is a poorly understood area: no one knows how to measure the health of an Interactive Business in a robust way, so we actually have to build many of these advanced metrics that transcend the very primitive state of metrics in the industry today. This is research and innovation in a live business environment!

The other challenges can be summarized by scale: both on data mining algorithm and on the management of data mining models. When you are responsible for generating thousands of predictive and classification data mining models, updating them daily, and then using them to produce predictions in real time, a huge challenge is to make sure that all these data mining models are updated, reading the correct information, and their outputs checked against what it takes to conduct data mining models and their notorious sensitivities to changes in the data or outliers. Many companies find it challenging to run a handful of models; we have to run and maintain thousands. This is a scale that is unfamiliar to most practitioners in our field and it requires systematic and product-like thinking -- not just analysis-oriented thinking.

Q3. GPS: After you left Microsoft, you founded digiMine (in 2000), which provided Data Warehousing, analysis and data mining applications via the Web. What was your biggest success there? Biggest failure ?

Usama Fayyad: We started DigiMine in March 2000. The idea was to evolve a hosted data warehousing and data mining on-demand service since most companies were unable to build their own systems effectively. Our biggest success is our unexpected traction with large Fortune 1000 companies. We quickly built up a big business with some of the world's largest companies in several verticals including Financial Services (e.g. American Express, GE), Manufacturers (e.g. DaimlerChrysler, Ford), Telcos (e.g. AT&T, T-Mobile), Technology Companies (e.g. Microsoft, Palm), Publishers (e.g. Wall Street Journal, CBS) and Retailers (e.g. Barnes & Noble, Nordstrom). The speed with which that distinguished list of clients grew was incredible and is something I am very proud of.

Looking back, we probably underestimated the costs and complexities of the business. While the business is

interesting, the fact that Business Intelligence and Data Mining are not well-defined and well-understood by the market, led us into situations where expectations of clients exceeded what technology could deliver. So even though we were deploying solutions that were more successful than what clients tried to build themselves, the lack of maturity in this business required us to build a growing Professional Services (Consulting) arm. While that business is very profitable, it is not scalable.

Ironically, this led to the ultimate creation of DMX Group which focused solely on the consulting business. The consulting also led to a very carefully chosen refocus of digiMine into one of the verticals we were operating in: ad targeting for online publishers and the evolution to Revenue Science today -- which is a strong player in their field. In the transition, I transitioned from being President & CEO of digiMine for over 3 years to becoming Chairman of Revenue Science and focused my energy on growing DMX Group.

DMX Group was thus a "spin-off" that absorbed the breadth of the mission of digiMine but was not pursuing the hosted business aspects. It was a wonderful situation because I had a book of business from day one, a profitable practice, and a company that was entirely owned by me (no venture capital needs). The allure of DMX Group was the pursuit of the toughest technical data mining challenges but in a context where the business users and executives had to clearly understand and value to work. In that sense, it was a unique effort.

Q4. GPS: After DigiMine, you founded DMX Group which was providing consulting services. What were some of the lessons you learned from DMX Group and the process of consulting?

Usama Fayyad: The biggest lesson I learned from DMX Group is that I love the consulting business -- which surprised me because I always thought of myself as a product guy. If you are a specialist in an area that is needed by your clients, and if you are delivering value, then there is no business that can teach you as much as the consulting business can. One of the hardest and most valuable lessons I learned was how to say no quickly, and how to carefully choose projects.

As a consulting company, you have to be careful not to take on too many projects, and you have to be very careful choosing your employees: the team IS the business. Recovering from one bad hire can ruin many other seemingly independent projects because when you have to recover, you have to draw on resources that are dedicated to other projects, and that is the start of the death spiral ...

I probably learned more from my time as President & CEO of DMX Group than I did in any other position I held before. The variety of projects and looking for common threads among what appear to be wildly differing client businesses is one of the most challenging and exciting pattern recognition problems I came across. What I learned from that experience is:

- much of what is done in the marketplace as one-off consulting work can actually be automated and productized.
The problem is that most consultants don't know how to think in terms of product and most product companies don't know enough consulting to get smart about what is needed!
- The market is ready for true data strategy consulting, but very few people out there truly understand what that means and how to provide it.

DMX Group proved to be an amazing growth story, and the only thing that could have stopped it for me was the completely unanticipated development with the Yahoo! acquisition and my new role at Yahoo!

Q5. GPS: Can you tell us about some Yahoo successes with data mining?

Usama Fayyad: Yahoo! is the first company to hire a Chief Data Officer - demonstrating that data is a true, strategic asset to the company. Our goal is to create value to consumers and marketers by delivering the

consumer-centric data platform & insight services that maximize user engagement and enable innovative marketing solutions. In a very short period of time, we have managed to have a significant impact on several Yahoo! products and services. The unfortunate thing here is that many of the contributions cannot be shared publicly (yet) because they involve competitive advantages or are specific to some of our advertising clients or products.

So rather than share the biggest successes, I'll share some of our more public ones:

Product integration: one of the examples that you see today on Yahoo! Mail is a visible result of data mining. One of the patterns we noticed in analyzing unexpected patterns in usage data was a strong correlation between people reading email and reading news, in the same session! When we shared this with the Yahoo! Mail product team, their first instinct was to test this effect: this was done by building a "news module" that highlights news headlines and showing this to a test group of consumers as part of the main Mail front page.

For a product like Mail, the critical business pain is to take new "light users", and turn them into "heavy users". If you do that, you reduce churn dramatically. Indeed, we showed from this test that churn in the weakest group was reduced by 40%. This led to the immediate development and launch of the News Module embedded in Yahoo! Mail's front page. Today, hundreds of millions of consumers see and use this product. I like this story because it highlights the quick reaction time of the product team who is obsessed with customer engagement. It also highlights the tremendous value lurking in the many, many patterns in the usage data.

Instant Messenger: we analyzed the drivers of the Instant Messenger usage, and out of the many factors, the data suggested that the most powerful factor to go after was to get users to grow their "buddy list" by at least 5 additional people. This led to a dedicated marketing campaign designed to achieve exactly this quantitative goal in order to increase the number of friends on your IM friends list and dramatically increase product usage.

Search box on Yahoo front page: A simple example had to do with discovering that on the Yahoo Front Page, centering the search box on the page (as opposed to having it be left-justified) would increase consumer usage. This led to better user engagement and there was no cost to Yahoo! to make the change. We discovered this by discovering the hidden pattern that showed that Netscape users tended to use search more than IE users, and by discovering that the only visible difference is the subtle position of the box! It was centered on Netscape browsers but left justified on IE browsers. A very unnoticeable difference, yet an important one. Who would figure that out???

Advanced targeting techniques: I would also list some of our advanced targeting techniques, such as discovering who is in the market for an automobile with high reliability, based on only anonymous browse data, as another big area of success for predictive data mining technology. This ability allows us to create very high-value audiences by understanding what they have in mind and biasing the advertising so it is more relevant to that audience: a win-win for consumers and advertisers; and of course a win for Yahoo! because reaching such purchase intender audiences commands a premium.

Q6. *GPS: What about Privacy and Data Mining at Yahoo?*

Usama Fayyad: Yahoo! was built around consumer trust and it has been our number one priority since day one. We would never do anything to compromise our users privacy and part of my role here at Yahoo! is to continue that tradition moving forward. In fact, one of my primary goals as Chief Data Officer, is to insure the "responsible use of data" in the words of our CEO, Terry Semel. Yahoo!, perhaps more than any other online consumer company, understands that the long-term survival and health of the business is built on consumer trust. In fact, one of the things that attracted me to Yahoo! was the company's extremely conservative position

towards privacy. Both are CFO and our COO always remind me of "long-term" value as opposed to what might appear to be a short-term bonanza. I truly respect that focus and it comes from the founders: Jerry Yang and David Filo.

As an example, when a competitor with a new mail product announced that they will target e-mail based on the content of the message, Yahoo! came out clearly saying that we will not do this. We consider the contents of your email as private information. Though funded by advertising, it represents an implicit assumption by the consumer that the content of the communication is private. This is a difference Yahoo! is extremely sensitive about: the expectation the consumer has as to whether they are in a "public space" or a "private space". The analogy I use is this: when you are in the shopping mall, you accept sales people approaching you or helping you find a product. When you are in your home, a sales person showing up in your living room would be an intrusion. As an officer representing Yahoo!, I will always err on the conservative side whenever faced with a judgment call. I find the public space/private space analogy very helpful in my reasoning about the subject.

Q7. GPS: Recently, there was a lot of [controversy whether Yahoo index size is bigger than Google](#) (also [here](#)). How big is Yahoo index and how does it compare with Google?

Usama Fayyad: In August 2005, Yahoo! Search reached a significant milestone -- our index provides access to over 20 billion items. This includes just over 19.2 billion web documents, 1.6 billion images, and over 50 million audio and video files. This happened to be a much larger index than our competitors were boasting to have and for some reason they reacted very negatively to the facts. The facts are there, and I am sure we will see the competitors soon reverse position and start announcing a much larger index themselves. We continue to believe that comprehensiveness (measured by index size) is only one dimension of the quality of a search engine. There are many other relevant metrics for relevance of search result, and we are obsessively focused on them: quantitatively and qualitatively. Finally, we don't comment on the index size of our competitors.

Q8. GPS: You also have a responsibility for Yahoo Research Labs -- can you tell us about some of the more exciting projects there?

Usama Fayyad: At Yahoo! Research, we have been working on some really interesting projects around prediction markets and Search technology. These are typically featured on [next.yahoo.com](#) Here are a couple of examples:

[Mindset](#) ([research.yahoo.com/research/data_analytics/mindset__intent-driven_search.shtml](#)) is a very innovative tool that analyses the content of web pages and constructs a score of "degree of commercialness" of a page. We then utilize this automatic categorization (which incidentally utilizes the world's fastest SVM [Support Vector Machine] learning algorithm -- also invented and built at Yahoo! Research), to allow the user to set a slider bar to reflect his/her intent.

If you are making a query and you are shopping for something, then commercial content in the search results is what you want. On the other hand, if you are researching something, you want to filter out commercial pages in favor of educational content pages. So moving the slider to the other end will re-order the search results, and many relevant pages that appear as rank 100 or 200 start appearing in the top 10. People who use this slider quickly get addicted to it (I am).

I encourage your readers to check it out. Note that this is only one dimension, and one can imagine many more that are important and many more sliders can be exposed: family-friendliness, personal, etc... Ultimately, if we know you enough, we can set these multitudes of sliders automatically for you.

[Tech Buzz Game](#) ([research.yahoo.com/research/foundations/tech_buzz_game.shtml](#)) is another example of an innovative and deep concept we launched. Back in March 2005, we launched an

innovative idea that is now becoming fashionable: a futures market in concepts! We had been working on that idea for over a year. The idea was to pick an area, in this case technology buzz as measured by search keyword activity, and provide a stock market for Tech Buzz Keywords.

We launched this as a game, but in the background we were doing the science: building the theory of what is happening and evaluating the predictive value of this marketplace in forecasting new trends in the technology market. The experiment was a huge success with over 20K players participating in the game we launched.

The winner actually used the optimal strategy derived by David Pennock, one of our Research Scientists who recently won the MIT Technology Review 35 Under 35 Award for this work. In fact, the winner beat randomized robots proving that the strategy works. This is an example of how we aim to build the sciences underpinning the Internet and Interactive Media.

Q9: *GPS: What do you do for fun, when you have time?*

Usama Fayyad: My most favorite activity is still sleep, whenever I can get it. :-)

With lots of hard work and an active life, I find deep sleep to be a wonderful pleasure -- probably because I rarely get enough of it. I very much enjoy skiing (both downhill and x-country), water skiing, and lots of swimming. I also enjoy playing chess, especially with my kids.

Q10: *GPS: What is your next ambition? Where would you like to be in 5 years?*

Usama Fayyad: In my experience, I have learned that the answer to that question almost always proves to be irrelevant. I can never envision correctly where I would be in 5 years.

If you asked me that question when I was a graduate student, I would have said I'd be a student forever.

Had you asked me that question when I was at NASA/Caltech Jet Propulsion Lab (JPL), I would have answered that writing a couple of deep books with some of the world-renowned scientists that I collaborated with on difficult analysis problems over massive data sets would have been my goal. Instead, I wound up at Microsoft learning all about platforms and stock options!

At Microsoft, I would have predicted that I'd be building the world's most innovative database platform. Instead, I ended up doing a startup.

Two years ago, I would have said I'd be doing startups forever and would have never guessed I'd be an officer in a public company, let alone the biggest internet company in the world! I am truly blessed to have thoroughly and passionately enjoyed all these roles, as I am my current one.

I still would like to get a couple of books out: one technical and one business!

Q11: *GPS: What is the most recent book you read and liked?*

Usama Fayyad: I have now made it a habit to read 2 books simultaneously: one business and one scientific or social.

The most recent are:

- *Fumbling the Future: How Xerox Invented, Then Ignored, the First Personal Computer*, by Douglas K. Smith, Robert C. Alexander and
- *Memoirs of the Second World War*, by Winston Churchill.

Churchill's book is amazing on how pertinent it is to today's world! How many lessons do we repeatedly miss...

Prabhakar Raghavan is now making me read *Modern Information Retrieval* (by Baeza-Yates and Rebeiro-Neto) which I am enjoying.

Q12: *GPS: Finally, what would you tell students who consider studying machine learning and data mining?*

Usama Fayyad: In my opinion, there is no technical field that will be more generally relevant in all sorts of businesses than statistical data mining. The world is truly drowning in data, and much like we spent the last few thousand years figuring out how to navigate and build structures in the physical world, the explorers and builders of tomorrow will be figuring out our journey and navigation technology in the digital data universe.

This is the Next Frontier, it is upon us, and we are truly lost in it. We can use all the help we can get.

Any smart young (and good) data miners out there: I think I can keep you very busy indeed!