DECISION STATS
BETTER DECISIONS === FASTER STATS

# Interview Dr Usama Fayyad Founder Open Insights LLC

Here is an interview with Dr Usama Fayyad, founder of Open Insights LLC (www.open-insights.com). Prior to this he was Yahoo's Chief Data Officer. In his prior role as Chief Data Officer of Yahoo! he built the data teams and infrastructure to manage the 25 terabytes of data per day that resulted from the company's operations.

**Ajay- Describe your career in science. How would you motivate young people today to take science careers rather than other careers**
**Dr Fayyad-** My career started out in science and engineering. My original plan was to be in research and to become a university professor. Indeed, my first few jobs were strictly in basic Research. After doing summer internships at place like GM Research Labs and JPL, my first full-time position was at the NASA – Jet Propulsion Laboratory, California Institute of Technology.

I started in research in Artificial Intelligence for autonomous monitoring and control and in Machine Learning and data mining. The first major success was with Caltech Astronomers on using machine learning classification techniques to automatically recognize objects in a large sky survey (POSS-II – the 2nd Palomar Observatory Sky Survey). The Survey consists of taking high resolution images of the entire northern sky. The images, when digitized, contain over 2 billion sky objects. The main problem is to recognize if an object is a star of galaxy. For "faint objects" – which constitute the majority of objects, this was an exceedingly hard problem that people wrestled with for 30 years. I was surprised how well the algorithms could do at solving it.

This was a real example of data sets where the dimensionality is so high that algorithms are better suited at solving it than humans – even well-trained astronomers. Our methods had over 94% accuracy on faint objects that no one could reliably classify before at better than 75% accuracy. This additional accuracy made all the difference in enabling all sort of new science, discoveries and theories about formation of large scale structure in the Universe.
The success of this work and its wide recognition in scientific and engineering communities let to the creation of a new group – I founded and managed the Machine Learning Systems group at JPL which went on to address hard problems in object recognition in scientific data – mostly from remote sensing instruments – like Magellan images of the planet Venus (we recognized and classified over a million small volcanoes on the planet in collaboration with geologists at Brown University) and Earth Observing System data, including Atmospherics and storm data.
At the time, Microsoft was interested in figuring out data mining applications in the corporate world and after a long recruiting cycle they got me to join the newly formed Microsoft Research as a Senior Researcher in late 1995. My work there focus on algorithms, database systems, and basic science issues in the newly formed field of Data Mining and Knowledge Discovery. We had just finished publishing a nice edited collection of chapters in a book that became very popular, and I had agreed to become the founding Editor-in-Chief of a brand new journal called: Data Mining and Knowledge Discovery. This journal today is the premier scientific journal in the field. My research work at Microsoft led to several applications – especially in databases. I founded the Data Mining & Exploration group a MSR and later a product group in SQL Server that built and shipped the first integrated data mining product in a large-scale commercial DBMS – SQL Server 2000 (analysis Services). We created extensions to the SQL language (that we called DMX) and tried to make data mining mainstream. I really enjoyed the life o doing basic research as well as having a real product group that built and shipped components in a major DBMS.
That's when I learned that the real challenging problems in the real-world where really not in data mining but in getting the data ready and available for analysis – Data Warehousing was a field littered with failures and data stores that were write-only (meaning data never came out!) — I used to call these Data Tombs at the time and I likened them to the pyramids in Ancient Egypt: great engineering feats to build, but really just tombs.

In 2000 I decided to leave the world of Research at Microsoft to do my first venture-backed start-up company – digiMine. The company wanted to solve the problem of managing the data and performing data mining and analysis over data sets, and we targeted a model of hosted data warehouses and mining applications as an ASP – one of the first Software as a Service (SaaS) firms in that arena. This began my transition from the world of research and science to business and technology. We focused on on-line data and web analytics since the data volumes their were about 10x the size of transactional databases and most companies did not know how to deal with all that data. The business grew fast and so did the company – reaching 120 employees in about 1 year.

After 3 years of doing high-growth start-up and raising some $50 million in venture capital for the company, I was beginning to feel the itch again to do technical work.
In June 2003, we had a chance to spin-off part of the business that was focused on difficult high-end data mining problems. This opportunity was exactly what I needed and we formed DMX Group as a spinoff company that had a solid business from its first day. At DMX Group I got to work on some of the hardest data mining problems in predicting sales of automobiles, churn of wireless users, financial scoring and credit risk analysis, and many related deep business Intelligence problems.

Our client list included many of the Fortune 500 companies. One of these clients was Yahoo! — After 6 months of working with Yahoo! As a client they decided to acquire DMX Group and use the people to build a serious data team for Yahoo! We negotiated a deal that got about half the employees into Yahoo! And we spun-off the rest of DMX Group to continue focusing on consulting work in data mining and BI. I thus became the industry's first Chief Data Officer.

The original plan was to spend 2 years or so to help Yahoo! Form the right data teams and build the data processing and targeting technology to deliver high value from its inventory of ads.
Yahoo! Proved to be a wonderful experience and I learned so much about the Internet. I also learned that even someone like me who worked on Internet data from the early days of MSN (in 1996) and who ran a web analytics firm still did not scratch the service on the depth of the area. I learned a lot about the Internet from Jerry Yaang (Yahoo! Co-founder) and much about advertising/media business from Dan Rosensweig (COO) and mTerry Semel (then CEO) and lots about technology management and strategic deal-making from Farzad (Zod) Nazem who was the CTO. As Executive VP at Yahoo!

I built one of the industry's largest and best data teams and we were able to to process over 25 terabytes of data per year and power several hundred million Dollars of new revenue for Yahoo! Resulting from these data systems. A year after joining Yahoo! I was asked to form a new Research Lab to study much of what we did not understand about the Internet. This was yet another return of basic research into my life. I founded Yahoo! Research to invent the new sciences of the Internet, and I wanted them to be focused on only 4 areas (the idea of focus came from my exposure to Caltech and its philosophy in picking few areas of excellence). The goal was the become the best research lab in the world in these new focused areas. Surprisingly we did it within 2 years. I hired Prabhakar Raghavan to run Research and he did a phenomenal job in building out the Research organization. The four areas we chose were: Search and information navigation, Community Systems, Micro-economics of the Web, and Computational Advertising. We were able to attract the top talent in the world to lead or work on these emerging areas. Yahoo! Research was a success in basic research but also in influencing product. The chief scientists for all the major areas of company products all came from Yahoo! Research and all owned the product development agenda and plans: Raghu Ramakrishnan (CS for Audience), Andrew Tomkins (CS for Search), Anrei Broder (CS for Monetization) and Preston McCaffee (CS for Marketplaces/Exchanges). I consider this an unprecendented achievement in the world of Research in general: excellence in basic research and huge impact on company products, all within 3-4 years.

I have recently left Yahoo! And started Open Insights (www.open-insights.com) to focus on data strategy and helping enterprises realize the value of data, develop the right data strategies, and create new business models. Sort of an 'outsourced version" of my Chief Data Officer job at Yahoo!

Finally, on my advice to young people: it is not just about science careers, I would call it engineering careers. My advice to any young person in fact, whether they plan to become a business person, a medical doctor, and artist, a lawyer, or a scientist – basic training in engineering and abstract problem solving will be a huge assets. Some of the best lawyers, doctors, and even CEO's started out with engineering training.

For those young people who want to become scientists, my advice is always look for real-world applications where the research can be conducted in their context. The reason for that is technical and sociological. From a technical perspective, the reality of an application and the fact that things have to work force a regiment of technical discipline and make sure that the new ideas are tested and challenged. Socially, working on a real application forces interactions with people who care about the problem and provides continuous feedback which is really crucial in guiding good work (even if scientists deny this, social pressure is a big factor) – it also ensures that your work will be relevant and will evolve in relevant directions. I always tell people who are seeking basic research: "some of the deepest fundamental science problems can often be found lurking in the most mundane of applications". So embrace applied work but always look for the abstract deep problems – that summarizes my advice.

**Ajay- What are the challenges of running data mining for a big big website.**

**Dr Fayyad-** There are many challenges. Most algorithms will not work due to scale. Also, most of the problems have an unusually high dimensionality – so simple tricks like sampling won't work. You need to be very clever on how to sample and how to reduce dimensionality by applying the right variable transformations.

The variety of problems is huge, and the fact that the Internet is evolving and growing rapidly, means that the problems are not fixed or stationary. A solution that works well today will likely fail in a few months – so you need to always innovate and always look at new approaches. Also, you need to build automated tools to help detect changes and address them as soon as they arise.

Problems with 1000 10,000 or millions of variables are very common in web challenges. Finally, whatever you do needs to work fast or else you will not be able to keep up with the data flux. Imagine falling behind on processing 25 Terabytes of data per day. If you fall behind by two days, you will never be able to catch up again! Not within any reasonable budget constraint. So you try never to go down.

**Ajay- What are the 5 most important things that the data miner should avoid in doing analysis.**

**Dr Fayyad-**I never thought about this in terms of top 5, but here are the big ones that come to mind, not necessarily in any order

a. ***The algorithms knows nothing about the data,*** and the knowledge of the domain is in the head of the domain experts. As I always say, **an ounce of knowledge is worth a ton of data** – so seek and model what the experts know or your results will look silly

b. Don't let an algorithm fish blindly when you have lots of data. Use what you know to reduce the dimensionality quickly. ***The curse of dimensionality is never to be under-estimated***

c. Resist the temptation to cheat: selecting training and test sets can easily fool you into thinking you have something that works. Test it honestly against new data, ***never "peek" at the test data – what you see will force you to cheat without knowing it.***

d. Business rules typically dominate data mining accuracy, so ***be sure to incorporate the business and legal constraints into your mining.***

e. I have never seen a large database in my life that came from a static distribution that was sampled independently. Real databases grow to be big through lots of systematic changes and biases, and they are collected over years from changing underlying distribution: ***segmentation is a pre-requisite to any analysis. Most algorithms assume that data is IID (independent and identically distributed)***

**Ajay- Do you think softwares like Hadoop and MapReduce will change the online database permanently. What further developments do you see in this area.**

**Dr Fayyad-** I think they will (and have) changed the landscape dramatically, but they do not address everything. Many problems lend themselves naturally to Map-Reduce and many new approaches are enabled by Map-Reduce. However, there are many problems where M-R does not do much. I see a lot of problems being addressed by a large grid nowadays when they don't need it. This is often a huge waste of computational resources. We need to learn how to deal with a mix of tools and platforms. I think M-R will be with us for a long time and will be a staple tool – but not a universal one.

**Ajay- I look forward to the day when I have just a low priced netbook and fast internet connection, and upload a Gigabyte of data and run advanced analytics on the browser. How far or soon do you think it is possible?**

**Dr Fayyad-** Well, I thnk the day is already here. In fact, much of our web search today is conducted exactly in that model. A lot of web analysis, and much of scientific analysis is done like this today.

**Ajay- Describe some of the conferences you are currently involved with and the research areas that excites you the most.**

**Dr Fayyad-** I am still very involved in knowledge discovery and data mining conferences (especially the KDD series), machine learning, some statistics, and some conferences on search and internet. Most exciting conferences for me are ones that cover a mix of topics but that address real problems. Examples include understanding how social networks evolve and behave, understanding dimensionality reductions (like random projections in very high-D spaces) and generally any work that gives us insight into why a particular technique works better and where the open challenges are.

**Ajay- What are the next breakthrough areas in data mining. Can we have a Google or Yahoo in fields of business intelligence as well given their huge market potential and uncertain ROI.**

**Dr Fayyad-** We already have some large and healthy businesses in BI and quite a huge industry in consulting. If you are asking particularly about the tools market then I think that market is very limited. The users of analysis tools are always going to be small in number. However, once the BI and Data Mining tools are embedded in vertical applications, then the number of users will be tremendous. That's where you will see success.

Consider the examples of Google or Yahoo! – and now Microsoft with BING search engine.  Search engines today would not be good without machine learning/data mining technology. In fact MLR (Machine Learned Ranking) is at the core of the ranking methodology that decides which search results bubble to the top of the list. The typical web query is 2.6 keywords long and has about a billion matches. What matters are the top 10. The function that determines these is a relevance ranking algrorithm that uses machine learning to tune a formula that considers hundreds or thousands of variables about each document. So in many ways, you have a great example of this technology being used by hundreds of millions of people every day – without knowing it!
Success will be in applications where the technology becomes invisible – much like the internal combustion engine in your car or the electric motor in your coffee grinder or power supply fan. I think once people start building verticalized solutions that embed data mining and BI, we will hit success. This already has happened in web search, in direct marketing, in advertising targeting, in credit scoring, in fraud detection, and so on…

**Ajay-  What do you do to relax. What publications would you recommend for staying up to date for the data mining people especially the younger analysts.**
**Dr Fayyad-** My favorite activity is sleep when I can get it J.  But more seriously, I enjoy reading books, playing chess, skiing (on water or snow – downhill or x-country), or any activities with my kids.  I swim a lot and that gives me much time to think and sort things out.
I think for keeping up with the technical advances in data mining: the KDD conferences, some of the applied analytics conferences, the WWW conferences, and the data mining journals. The ACM SIGKDD publishes a nice newsletter called SIGKDD explorations. It is free with a very low membership fee and it has a lot of announcements and survey papers on new topics and important areas (www.kdd.org).  Also, a good list to keep up with is an email list called KDNuggets edited by Gregory Piatetsky-Shapiro.

**Biography (www.fayyad.com/usama )-**

Usama Fayyad founded Open Insights (www.open-insights.com) to deliver on the vision of bridging the gap between data and insights and to help companies develop strategies and solutions not only to turn data into working business assets, but to turn the insights available from the growing amounts of data into critical components of an enterprise's strategy for approaching markets, dealing with competitors, and acquire and retain customers.

In his prior role as Chief Data Officer of Yahoo! he built the data teams and infrastructure to manage the 25 terabytes of data per day that resulted from the company's operations. He also built up the targeting systems and the data strategy for how to utilize data to enhance revenue and to create new revenue sources for the company.

In addition, he was the founding executive for Yahoo! Research, a scientific research organization that became the top research place in the world working on inventing the new sciences of the Internet.

DECISIONSTATS.COM