Anand Rajaraman Speaks Out on Startups and Social Data

by Marianne Winslett and Vanessa Braganholo

Anand Rajaraman
http://anand.typepad.com/datawocky/anand-rajaraman.html

Welcome to ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we are in Phoenix, site of the 2012 SIGMOD and PODS conference. I have here with me Anand Rajaraman, who is an entrepreneur from the database research community. Anand was a cofounder of the data integration company Junglee, the semantic search company Kosmix, and the venture capital fund Cambrian Ventures. After Amazon acquired Junglee, Anand served as Director of Technology for Amazon.com. After Walmart acquired Kosmix, Anand became the senior vice president and co-head of @WalmartLabs. After leaving Walmart in 2012, Anand continues to invest in, mentor, and advise several Silicon Valley startups. He has a VLDB 10 Year Best Paper Award[1] and a SIGMOD Test of Time Award[2]. His PhD is from Stanford University. So, Anand, welcome!

Your two 10-year best paper awards are both for papers that you wrote in 1996, which was also the last year that you published a research paper! What happened?

[1] Querying Heterogeneous Information Sources using Source Descriptions. Alon Halevy, Anand Rajaraman, and Joann J. Ordille, 1996.
[2] Implementing Data Cubes Efficiently. Venky Harinarayan, Anand Rajaraman, and Jeffrey Ullman, 1996.

Sometimes people have these years they call *annus mirabilis* (miraculous years). So 1996 was my *annus mirabilis*. In 1996 I did two streams of research, which ended up with these best paper awards, one at SIGMOD and one at VLDB. But at the same time, I also came up with the idea for my first company, Junglee, together with some other students at Stanford. That ended up being the year that I took a leave of absence from the PhD program at Stanford to start my first company, Junglee, and after that, I never published any refereed research paper. (I misspoke: it turns out I do have a couple of research papers after 1996. But they have been sporadic).

How did your advisor feel about that?

My advisor was Jeff Ullman at Stanford, and you know, I really credit Jeff with everything that has happened to me since that time. We'd been doing all this research work on how to do data integration by combining all these enterprise databases; it was a big project at Stanford called TSIMMIS. A bunch of us had this idea that you could take some of these ideas, but not the exact technologies, and apply them to something new that was coming up called the World Wide Web. As we thought about it, it became clear to us that the right way to pursue this was not as a research project, but as a company, as a Silicon Valley startup. And we were also highly inspired by meeting the Yahoo! founders, who at around the same time had left Stanford and started a company as well. So I was kind of in two minds. Should I leave Stanford and start this company? Here I was, I had just published these papers, it looked like my research career was finally going to take off, and at the same time, I had this idea for a company, which I thought was a truly interesting idea that could change the world. So what should I do? So, I spoke to the person who I thought had the most insight and this happened to be Jeff Ullman. And Jeff said: "You know what, if you truly believe in this idea, then go make the company happen. Building a great product that many, many people use is far more impactful than writing a thesis, so just go ahead and make it happen".

It's never been so easy to collect data than it has been now, so never take the data as a given.

If I were cynical, I should ask you if he had shares in the company.

Well, that's a good question! At the time Jeff gave me the advice, he had absolutely no shares in the company or any interest in the company. Later on, as it turned out, as we got further along in the company, we actually added Jeff to the Board of Directors at Junglee, but this was much later. You know, maybe several months later after this conversation.

So, it seems that, in the intro, I should have said, "If Anand had a PhD, it would have been from Stanford?"

Actually, I do have a PhD, and it is from Stanford. But here's the story. I took a leave of absence in 1996, as I said, to start this company, Junglee. In 1998, Amazon.com acquired Junglee, so I went to Amazon and did a bunch of interesting things. But most importantly, I worked around how to get third party merchants selling on Amazon. The Amazon Marketplace that you see today is what the Junglee team did at Amazon. Then around 2000, I had this feeling that there was something incomplete, I needed to get closure on this whole PhD thing. So I came back to

Stanford in the year 2000, spent a year, and wrote up my thesis. But here's one thing that is still a sore point. When I came back, Jeff Ullman told me that he wouldn't actually give me a scholarship to finish my PhD. He said, "You know you've got to pay your own fees". So I did that.

That's so inconsiderate!

(Anand laughs)

I would say that among the successful founders of companies, you'd be in the minority in the fact that you came back and finished that degree. What was motivating you that other people didn't feel? Like the Google guys, they didn't come back and finish.

I guess they were far more successful than I was. If you look at the people who actually started companies that kind of took off, and then they stayed for a long time at those companies, they actually haven't ended up coming back to finish their PhDs. In my case, two things happened. One is that my company got acquired within a relatively short time after I started it, and so my research was still fresh, so I could come back and complete my thesis, so that was good. And the other interesting thing that happened was that I really wanted to finish this, so I just did it.

What will e-commerce be like five years from now?

Do you remember the time before e-commerce when you actually had to go to the stores to shop? Then e-commerce sort of happened in the early 90's, and there was a huge change in the way people shopped, right? So the way we shopped changed fundamentally with e-commerce, and a fundamental change, as fundamental as that change, is just happening now to the way we shop. That's kind of driven by two factors. One is social, and the other is mobile. These days, more and more shoppers are carrying smartphones, and they use these smart phones, not necessarily to make phone calls, or to check the weather, but also to compare prices, and find where to buy products. We spend more and more time on social media, and what our friends say about what products they buy and so on deeply influences our purchase behavior. So because of social and mobile, e-commerce is going to change fundamentally, and it is going to be as big a transformation as e-commerce was.

There are two distinct worlds today. There is the world of e-commerce, and there is the world of retail commerce, where you shop offline. Because of mobile, these two worlds are going to merge together. The distinction between what we call e-commerce and what we call retail is going to go away. And it is going to be one seamless customer experience. Customers won't care or won't even know sometimes that they are shopping online or offline. For example, you could go to a shop, see that the product that you want is out of stock, and order it online and have it shipped to your home. Do you call that retail or do you call that e-commerce, right? Or you could go online and have a product shipped to your nearest store and you can go pick it up there, now is that retail, or is it e-commerce? So all kinds of interesting combinations will come into play that will completely blur the line between e-commerce and retail, and this whole category of e-commerce is going to go away, there is just going to be commerce.

Now I'm confused, because the two examples you gave already exist. For example, if you shop at Talbot's and what you want isn't there, they do that and have it shipped, and the reverse direction also works. So where's the new angle?

Right, so, all these are trends that are starting to happen. There are early experiments in these things by a few online retailers. But these will become the new reality over time. The mobile will be an incredibly important part of the retail experience. Today, there are a lot of things, for example, when we shop online, there is a lot of stuff that we take for granted. For example, we read reviews, and we see what other people have said, and so on. Yet, when we go into a store, we have none of those things. We just see shelves of products, right? So if you think about the first generation of e-commerce, it was all about taking the products that were in the store, and bringing them to the web. The second generation, now of commerce, is going to be taking all the information about products that's online and bringing them into the store through the mobile phone, and then using your social identity to connect the rest of your behavior with your shopping behavior.

One of the most quoted examples from e-commerce is Amazon's feature of what you should read based on what other people similar to you have read. I was at Amazon at the time when they launched that, and it is truly a brilliant feature. If you think about it, the only information that Amazon, or any other e-commerce site has access to right now is your shopping behavior on that site. Yet, there is so much of our life beyond what we spend at any one website. And that behavior has more and more been captured in social media streams like Facebook and Twitter. So if you can combine the information that's in Facebook and Twitter about us together with the information that the retailer has, and deliver all those recommendations and the better search experience through mobile, that's going to be truly revolutionary.

[…] the most successful uses of big data […] use all the data to answer the questions. They don't ever throw away the data, they are kind of "model light and data rich".

So speaking as an introvert here, how will that make my life better? Except, for example, maybe if I am buying a car, or some other mega-purchase? If I'm buying socks, how's that going to make my life better?

Well, I'll give you an example that happened to me: I work out and I run, and my feet blister easily, so I needed to find socks that would not blister. So I asked my friends, and they told me, "we also run, and these are the socks to buy". Now, it would be nice when you are in a store to ask your friends right from there: "which socks should I pick up"? These are the kinds of things that you might find interesting, for instance. How do you connect with your friends when you are in the store, how do you leverage recommendations, how do you leverage the wisdom of your friends as well as the whole community when you are shopping, in a better way? How do you get personalized recommendations? For example, let's say you're traveling somewhere, and you just happened to go into a store that has the right

guidebook for where you are traveling. Well it might be interesting for you to get an alert to your phone saying, "hey, you know, the product you are looking for is right here".

Speaking of Amazon, where did Amazon's Mechanical Turk come from?

That's an interesting story. I told you that I left Amazon around the year 2000 and came back to Stanford to complete my PhD. At the same time as I was working on my PhD, together with another Junglee cofounder from Stanford, Venky Harinarayan, we started what we called an idea incubator, called Cambrian Explosion, which is an arm of Cambrian Ventures, a venture capital firm. With Cambrian Explosion, we were interested in coming up with new ideas that could potentially become interesting business. And one of the ideas that we were playing around with at the time was this idea of how we combine humans and machines to complete interesting tasks. What we observed (this was around the year 2000) is that computers are great at doing some things, but there are some things that computers are terrible at doing that humans do effortlessly, like image recognition and things like this. So we thought that if we could combine humans and computers, and create what we call hybrid human-machine computation, we could solve a wider area of problems. So we sort of came up with this idea, and found a couple of entrepreneurs, who were in fact willing to take this idea forward. We wrote up a patent called Hybrid Human-Machine Computation, filed it in 2000, and started a company to take the idea forward.

Our idea at the time was we could build software that would enable companies to write systems combining humans and machines in interesting ways. So we had these two founders of this company who were going to do this, and they were talking to a whole bunch of potential customers to see whether they could use humans and machines together to solve interesting problems and so on, and we were getting some interest. But, as it turns out, just around this time, 9/11 happened, and companies stopped trying to do new things. Kind of, the bottom fell out of innovation around that time. And so, it sort of became apparent to us that this company that we had started around hybrid human-machine computation wasn't going anywhere. The two entrepreneurs with whom we were working on that came up with a different idea they got more passionate about.

So here we were: we were sitting on this idea that we thought had potential, but we had no people to take it forward. This was when we had a chat with Jeff Bezos. Incidentally, when we left Amazon, Jeff Bezos wanted to stay engaged with us, and was in fact, the biggest investor in our venture capital firm, Cambrian Ventures. When we told him about this idea about hybrid human-machine computation, he got incredibly excited. He said "look, I'd like to take this idea forward. Why don't you guys sell me this patent?" So we sold him the patent on hybrid human-machine computation, and that became the basis for Amazon Mechanical Turk. So the name "Amazon Mechanical Turk" is entirely Jeff Bezos's. We had nothing to do with it. Jeff's genius in this was to take that idea, and combine it with the idea of a market place. It was sort of saying that you could have this marketplace of humans, and you could create these tasks and you could put it out there, so that was his thinking. And then Amazon executed very well on the idea, and it became quite successful. So that was our contribution to Amazon Mechanical Turk.

You have claimed that more data almost always beats better algorithms. Why is that?

You know, we live in a world where there's more and more digital data that's being created. And usually people pull out statistics about how data is growing at 50% year over year. But my favorite quote on this is from Eric Schmidt who said that every 2 days now, we create as much data as was created from the dawn of civilization until 2003. That's a huge amount of digital data that's being created. When I think about how to solve difficult problems, I always think about how do I leverage all this data to solve that difficult problem. Now, if you think about data driven applications today, most of them follow a certain paradigm. You sort of create your favorite machine learning model, whether that's support vector machines, or regression, or whatever it is, and then you use all this big data as training data to train this algorithm. Then, once you have the algorithm, which is the trained model, the parameterized model, you through away all the data, and then you just ask the questions directly to the model. What a waste! Because you've thrown away all this data, and you've tried to capture everything, all the intelligence, in this model.

It's a well-known phenomenon that as you keep throwing more and more training data at a given machine learning model, the precision-recall performance of the model saturates at a certain point. At this point, if you want to get better at prediction, the only thing you can do is to make the model more complex by adding more features. But the problem is, the more complex you make the model, the more likely you are to be wrong. Just because the world is a fundamentally complex and a changing place, and all this complexity in the model probably means the world has diverged away from the model over time. So, if I think of the most successful uses of big data, like Amazon's recommendations, which is an example of collaborative filtering, or Google search, which I think is the best data driven application out there, both of these applications use all the data to answer the questions. They don't ever throw away the data; they are kind of "model light and data rich". I think that that's the right paradigm to think about how to leverage big data. Never throw it away once you've trained a model, keep it around all the time and use all of it to do every task, and come up with light thin models that are like icing on top of the data rather than try to replace the data by a model.

… we live in a world of big data, and there's never been a better time for startups around the idea of data.

What about things like smoothing that help you model the data that doesn't yet exist.

That is a very good point. One of the things that you run into, especially with high dimensional data sets, is the sparsity problem. When you try to find nearest neighbors in high dimensional data, if you have a certain number of data points, and the dimensionality of your data cell increases, then they, on average, get further and further away, so finding nearest neighbors becomes harder and harder. In my experience, one of the best ways I've found of dealing with this is through dimensionality reduction, to the extent possible, and then to just keep getting more data. Throwing more and more data into this mix. I think smoothing is a way of compensating for the lack of data, but we are transitioning from a data poor world into a data rich world. So while compensating for lack of data is interesting, I think we should be thinking about how to leverage all this extra data that's coming online.

In Google's case, don't they use hundreds of features, isn't that very high dimensional already?

I am not entirely familiar with the details of the technology behind Google search. I am sure they use hundreds and hundreds of features, but the key is that the data is fundamentally the lever, and the algorithms are the fulcrums, it's not the other way around. They don't talk about training data, the index is not the training data, the index is the data, and it answers every question.

Well, how can a database researcher know when the payoff is in collecting more data, and when to focus on modeling the part they haven't seen? I mean, the fatter the tail, the more you'll never see, to how do you know whether you should work on a model or work on getting more data?

I think it depends on the problem you're solving. So there's definitely no "one size fits all". But the one thing that I would say is that it's never been easier to get more data. So the way I like to phrase it, is don't ever take the data as a given when approaching a problem. I teach students in the data mining class at Stanford as well, and many students tend to approach the data as a given. The data is never a given. You can always collect more data. It's never been more easy to collect data than it has been now, so never take the data as a given. Always look for complimentary data sets, or additional data sets. I think time spent doing that is usually more rewarding than time spent designing more complex algorithms.

We talked a lot about startups. Do you have any advice for database researchers who would like to have a startup?

Well, they should just come talk to me! Seriously, what I mean is, you know, we live in a world of big data, and there's never been a better time for startups around the idea of data. If there is any database researcher who wants to start a company, the time is now, there's no time like the present. And I am happy to sort of talk to any of them, and help figure out how to take it forward. But, I think there are huge opportunities in the area, specifically around big data, in the infrastructure layer. And there is another trend that I'm sort of starting to see merge around fast data, which is data that's big but data that's moving faster and its real time. For this, there are opportunities in the infrastructure layer, in the algorithm layer and in the application layer, so there's huge opportunities, and now's a great time to be doing startups.

Well, these startups are usually West coast US, what about for all the people in our audience who live in other parts of the world?

Move to Silicon Valley! It worked for Mark Zuckerberg.

So, proximity is key?

Well, I think it's not necessarily about proximity. I think Silicon Valley has a great ecosystem that helps startups succeed.

What about Bangalore?

You know, I do see some interesting startups in Bangalore, I was just in Bangalore about a month ago, and I met with some very interesting startups there.

Beijing?

I have not been to Beijing, so it is hard for me to tell.

Maybe in time, there will be places other than Silicon Valley.

It is quite possible. And I know Silicon Alley in the New York area is immerging as an interesting startup hub as well. But I've found there is no place to beat Silicon Valley.

Right now you are at @WalmartLabs. What's that extra "at" there for?

Sure, you know how on Twitter and on Facebook when you want to address someone, you put an @ in front of their name? It's sort of a handle. So we built @WalmartLabs in the same sense, because @WalmartLabs is all about combining social into commerce, so we thought we'd sort of make a point by putting the @ in front of our name. And that also happens to be our handle, so that you might want to follow that handle on Twitter.

What are you guys doing with social media?

We are doing experiments on how is commerce best done using social media. For example, one of the experiments that we've done is something called Shopycat[3]. This is a Facebook App that we launched for the last holiday season, and what this Facebook App does is that it sort of takes the pain out of gift giving. So in the holiday season, we all want to give gifts. We have so many people in our lives, and we want to give them thoughtful gifts, not just a gift card. You want to give a thoughtful gift, and a gift that you think they are interested in. But how do we keep track of all that? Well it so happens that we tell our friends on Facebook everything that we're doing. And there's a set of information in there to figure out your hobbies, your interests, and so on. So what Shopycat does is for each of your friends, it figures out what their hobbies and interests are, combines them with a giant gifting catalog, and comes up with interesting gift suggestions for each of them. For example, you might find out that one of your friends is into hiking and the other is into running, and you can give them a different pair of shoes, hiking shoes or running shoes. And if you have a younger relative, you can find out that she's into the Hunger Games, and you can get her some Hunger Games memorabilia.

I think this data about human beings […] is going to create a revolution that's as fundamental or more fundamental than the industrial revolution.

Is it true that you were offered a chance to buy Google and turned it down?

[3] https://www.facebook.com/Shopycat

That's an interesting story. Remember, this is back in the year 1998, when the company that we had cofounded, Junglee, was in the process of being acquired by Amazon. So we had sort of agreed to be acquired by Amazon, but the deal had not closed yet, and around the same time, Sergey and Larry were getting started with Google. But they hadn't quite figured out how to make it a big company or whether it was going to be a big company even at that point in time – that was back in 1998. So it so happened that Sergey's advisor is also Jeff Ullman, who's my advisor. And Jeff connected us to Sergey and Larry and then he mentioned they were trying to figure out what to do, and perhaps Junglee might be interested in acquiring the company. The search technology at that time was relevant to what we were doing, you know, we were doing product search, they had some web search, maybe there was some synergy and so on. So it seemed very interesting to us. The problem for us is that we were in the process of being acquired by Amazon, so when you are in the process of being acquired yourself, you can hardly go around acquiring other companies. So that is why we couldn't do it at that time. Interestingly, there was another incident in the year 2000, or maybe in the year 1999, when we were at Amazon and we were talking to Jeff Bezos, and we were seeing Google starting to take off. This was the early days, but we could see the potential, and we convinced Jeff Bezos that Amazon should acquire Google. So Jeff Bezos sent me down, together with a couple other people to visit Google headquarters, which were in Palo Alto, and try to buy them. We were authorized to offer up to 300 Million dollars to buy Google, but when we met Sergey and Larry, they wouldn't budge for anything less than a billion dollars. So that didn't happen either.

Has being married to a fashion designer improved your fashion sense?

The distinction between what we call e-commerce and what we call retail is going to go away.

What do you think? (He laughs.) No, seriously, it is great fun being married to a fashion designer, because that's very remote from database technology, as you can imagine. And it gives you a different perspective on life. You know, I especially like the fashion shows and being able to go back stage during the fashion shows, and all that stuff. It's just different. And the parties are a lot more fun!

Maybe you can get some invitations for some members of our community! This cross-fertilization is probably good.

I would be happy to!

If you magically had enough extra time to do one additional thing at work that you are not doing now, what would it be?

You know, I would personally get my hands dirty and play with big data more than I am. At @WalmartLabs, we've set up this giant, big fast data cluster, with many many nodes. There is lots of interesting data analysis going on that combine Walmart's data with Twitter and Facebook and so on. Fascinating stuff. And I wish I had the time to actually go do some of that myself. Unfortunately, when you get to a point when you are managing a large organization like

this, you tend to play more of an advisory role to the people who are actually doing the really fun stuff. So I wish I had more time to do some of that stuff myself.

If you could change one thing about yourself as a computer science researcher, what would it be?

You know the one thing that I would love to do more is to actually spend more time being a computer science researcher. My career, as you mentioned, has been in startups and in venture capital. When you are a venture capitalist it turns out you can actually spend a lot of time doing interesting stuff because there's not much to do otherwise. But when you are running a startup, or when you are working for a company, you don't have that much time to do real research. I try to spend as much time as possible at Stanford, in fact, I teach a class there on data mining. And I dearly love interacting with students, and I wish I could do more of that, and do more computer science research, and come to more conferences like SIGMOD and interact with the great people here. You know, I find it so refreshing to be able to do that, I wish I had more time to do that.

If you had that time, would you work on big data and social media, or would you pick a different topic, something different from your current day job?

I definitely think social media and social data is something really huge. The way I think about this is the following. If you go back a few hundred years, to the 16th century, there was this guy called Tycho Brahe and he observed the heavens and he jotted down the positions of the moons of Jupiter and all these things in a big book, and that I think was the first real database. And it lead to wonderful things, like Kepler's laws of planetary motion, and Newton's equations, and it lead, indirectly, to the industrial revolution, which changed the way we live. Now, if you think about all the advances that have happened in Physics, and in various other fields, that have actually been transformative for the world, many of them have started from physical observations of phenomena that are in the cosmos and all around us. What's been lacking until now, when we wonder about the laws of cosmos, and the laws of physics, we lack a fundamental understanding about human beings and human societies, and what makes us tick. And for the first time in our lives, due to social data, we have more data about human beings than ever before. I think this data about human beings is actually more valuable than the data about the cosmos, and it is going to create a revolution that's as fundamental or more fundamental than the industrial revolution. And if I can in some way play a small part in that, that's what would give me the greatest pleasure.

That is really exciting. So we have a lot of young readers who may be reading what you say and inspired by it, and then the next question in their mind would be how do I get access to this incredible data set? So how do they, how can they do that if they don't work for Facebook, Twitter, etc.?

That's right. One of the nice things about social platforms is that, for example, Facebook has a platform where you can create a Facebook App, and if you can get people to install your Facebook App, then you get access to their data. So I would encourage people to start creating Facebook Apps that are useful for people to use, and then that gives them access to data of people, which they can then use. So that is one way of gathering data on Facebook. On Twitter,

you can license Twitter data, or you can license them on relatively cheap terms. And I would highly encourage pretty much every university to go get cracking on that.

Well, thanks very much for talking with me today.

Thank you, it has been a pleasure, Marianne.