Jeff Vitter Speaks Out on being a Southerner, Duties of a Dean, and More

by Marianne Winslett and Vanessa Braganholo

Jeffrey S. Vitter
http://provost.ku.edu/jsv

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today I am at Purdue University. I have here with me Jeff Vitter, who is the Frederick L. Hovde Dean of the College of Science[1]*. Before coming to Purdue, Jeff was on the faculty of Duke and Brown for many years, and he served as the chairman of the Department of Computer Science at Duke. Jeff's research interest lies in algorithms, especially in the areas of external memory algorithms and compression. Jeff is an ACM Fellow, IEEE Fellow, and Guggenheim Foundation Fellow*[2]*. He is on the board of directors of the Computing Research Association and is the former chair of ACM SIGACT. His PhD is from Stanford. So, Jeff, welcome!*

Great! Thanks for having me, Marianne.

[1] This interview was conducted in 2008. Today, Jeff Vitter is the provost and executive vice chancellor and the Roy A. Roberts Distinguished Professor at the University of Kansas.

[2] In 2009, Jeff was elected as a Fellow of the American Association for the Advancement of Science (AAAS).

Jeff, what was it like working with Don Knuth at Stanford?

Don is just an incredible human being. You know, he is really probably more responsible than any other person for the founding of computer science as an academic discipline. So just getting his insights was really tremendous. Professionally, what really impacted me was his sense of importance of theory and practice, and how it's vital to have a deep understanding of them both in order to excel at either. It was a little intimidating because he had just started TeX, and he really wasn't taking students, and one day I went to see him and I told him I had solved this problem and I thought this other one might be interesting to look at, just to see what he thought. And he said, "Well, if you do that, that would make a great thesis. And, by the way, you should plan to do this here, then this then, and graduate at this time," which was in three years from when I got to Stanford. I didn't dare question this. I just plowed ahead and did it. And I remember going through my thesis near the end of my third year, getting ready to finish, and Don looks at me and says, "You know, you really did quite a bit here, in an amazingly short amount of time. Why did you do it so quickly?" And I am sitting there after having worked so hard, and I was about to say, " 'Cause you told me to!" (Laughing.) But it was just a great experience. He was the most remarkable academic I have ever met.

So, what was that thesis on?

It was on Coalesced Hashing, as it's called. It is a hashing method that optimizes the way it uses storage in order to get the absolute best in search time. I have adopted the name "Coalesced" for some our projects here in the College of Science.

Jeff, most of your research is on algorithms for massive data sets. But your papers mainly appear in theory-oriented venues like Algorithmica and FOCS (Foundations of Computer Science), rather than SIGMOD, VLDB, and ICDE. So are you a theory guy, a database guy, or a database theory guy?

Yes (grinning).

To follow up on what I learned from Don, I think the most important things are this blending of theory and practice, so that is what I try to instill in my students. I really try to cover both of those communities. I have had some great students who have gone on in the systems arena, but because they have such a strong theory background and can appreciate the elegance and essence of what the techniques they are working on are all about, I think that really brings a scalability that makes what they do in systems work out. I have had students like Mark Nodine, Paul Howard, Dzung Hoang, Darren Vengroff, Lipyeow Lim, Tavi Procopiuc, Rakesh Barve, and Min Wang. They are incredible systems implementers, but they are also fundamentally very strong algorithmic students. I think that is part of the reason they are so good in systems.

What is the relationship between compression and database query optimization?

> ***Academic administration […] is really computer science on a grander scale. It is problem solving, or to put it more positively, it is finding solutions***

Historically, histograms are used a lot in order to summarize what's happened in the past to guide decisions for query execution or whatever. And my interests in this field, I really have a variety of different interests, and that is really what drives me as a researcher. One of my grad students, Min Wang, and I were working in the area of looking at compression because I was looking at compression from a variety of fronts, and along with Yossi Matias, we collaborated on applying wavelets. It was really the first time wavelets were used in the database community. It was used in a way to really be a novel form of histogram; capturing data in a fundamentally more efficient way, more effective way. So we worked out a lot of algorithmic aspects, it was very effective for doing this kind of query estimation we are talking about, or doing approximate answers, if you are in OLAP-type query situations. That has led to a lot of other work where wavelets have proven to be very effective. There have been great results by some on how to get provable bound estimates through wavelets. So, that has been a very exciting thing. But you know, the goal of all of these areas is really prediction. If you can do a better job of predicting what will happen in the future, you are going to be able to have a more effective system, more efficient or whatever.

Prediction is really nothing more than learning. It is trying to understand what will happen. That has driven a lot of my fundamental research. So, to give you an example, let's take a learning problem, which is the same as prediction, trying to learn what an elephant is. Suppose I want to teach you what an elephant looks like. This is actually very relevant in this U.S. Presidential election year, cause a lot of people are trying to understand what elephants look like. So here is the problem, I am going to give you a bunch of photos of animals, and I am going tell you for each one if it is an elephant or if it is not an elephant. And after a while, hopefully if you are a good learner, you will be able to know what an elephant is. So if I give you a new picture, that you haven't seen before, you will be able to tell if it is an elephant or not correctly. So, in the computational learning theory area, there is a domain called "PAC learning,"[3] where you can actually prove that learning is the same thing as data compression in the intuitive sense that if you as a learner do nothing more than memorize the pictures I showed you, you are going to have no chance of then classifying this new picture. But if instead, you have compressed what you have seen into a few basic rules, like elephants are grey, they are big, they have a trunk, they do not have wings, and things like that, then you will have no trouble classifying the

[3] PAC learning stands for probably approximately correct learning.

new picture as to whether it is an elephant. And that is really the essence of this relationship.

So, we were looking at a variety of problems, one of them was prefetching. Prefetching is a job where you have a bunch of accesses to a disk in the past, and now based on those you want to predict what are you going to access in the future so you can prefetch it into memory and have it ready for when you are going to access it, and avoid a costly page fault. So we applied a data compression method because of this intuition that compression is really prediction. We applied a data compression method to the sequence of numbers, which are page accesses, and in the bowels of the method, the Lempel-Ziv method, was a prediction for what the next page reference would likely be. We used that, and we showed that actually it allowed us to boost the hit rate from 20% in many applications, up to 70%, so it was very effective. And it has a really nice mathematical foundation. So prediction and compression come into play in a lot of instances. In image databases, it's the key for storing images so that you can search for them based on similarity. And of course, any time you have compressed data, it will often be stored in faster areas in the memory hierarchy, and then it makes it more efficient.

You wrote the book — literally[4] — on external memory algorithms. What are they, and how do they relate to databases?

It all goes back to a model of memory hierarchies, or what we call a parallel disk model, where in a simple setting, we have a computer with an internal memory and data are simply too large to fit in the internal memory, so we store it on disk. And this is a standard database set-up. Because disk drives are these physical rotating media where it takes milliseconds to get to data, but once you get to data, you can get adjacent data very quickly, the result is that data are typically transferred in blocks because that amortizes the cost of the high latency just to go to the data. One of the main goals of external memory algorithms is to minimize the number I/O transfers. And I/O is transferred in large blocks of data, so the main parameters of the model are the size of the transferred block, the size of the internal memory, and then basically that's it, the problem size itself. And the goal is to design an algorithm that uses locality in a fundamental way, so that data are transferred in blocks, and when you want data, you want a block of data, you don't want data from random locations, because if you do things effectively, you can speed up computations by a factor of 100 or 1,000 because of this block mechanism. So to give you an example, we applied this in a domain at Duke in collaboration with some folks in the School of the Environment. Lars Arge and I and students and collaborators in the School of the Environment worked on methods for determining, when rain falls, where it will go. So, what will the watershed be? Where will the flooding occur? This is very

[4] J. S. Vitter. *Algorithms and Data Structures for External Memory*, Series on Foundations and Trends in Theoretical Computer Science, Now Publishers, Hanover, MA, 2008. Also published as Volume 2, Issue 4 of *Foundations and Trends in Theoretical Computer Science*.

important in North Carolina. So we took satellite data and other imaging methods of regions like the Appalachians, and using so-called conventional techniques, such as ArcInfo, these calculations could take several days. There would be calculations that could not be run at all. Using newly-designed algorithms that focus on block transfer, we were able to reduce the running time from days to hours, or when they couldn't even be computed at all, we could do them in just a few hours. So it can make a really big difference, especially because data are just expanding at a crazy rate.

You are a relatively recent transplant from the east coast to the Midwest. What do you think of life in the Midwest?

I grew up in the south, went to grad school in California, and then I was at Brown and Duke on the east coast. But I did go to Notre Dame as an undergrad, so I have strong roots in Indiana. I am happy to say that being two hours south makes a big difference in temperature. It is a lot warmer and more moderate here. The main thing about Indiana is it is a great family environment. West Lafayette in the last 10 years has gotten some really wonderful restaurants, culture opportunities; in fact, there is a New Orleans restaurant that just opened a couple of months ago, and the owner and chef is a high school classmate of my brother Mark, so it is really good. It's a great place to live. And the students here are, with their Midwestern ethic, just very hard workers. They are wonderful to work with.

[…] in the arena of the life sciences and biology, there are great opportunities that put databases at the forward.

Some people think that CS researchers who aren't on the east or west coasts must be quite isolated. Have you found that to be true?

It is a perception that is challenging at recruiting time, but when you show the candidates all that is going on, all that we have at Purdue, it is really quite remarkable. In databases, with this community, we have an incredible group. We have Ahmed Elmagarmid, Walid Aref, Elisa Bertino, Chris Clifton. It's a great group. Ahmed is actually the head of the Cyber Center, which integrates IT research across the entire University. In information security, we have what I think is the best group anywhere. 25% of all of the information security PhDs in the entire country come out of Purdue and our CERIAS Center. Mike Atallah and Gene Spafford are just renowned in that area. We have terrific systems people, whether it is in networking, distributed systems, or programming languages, operating systems, graphics and visualization, software engineering. It is really a strong group. So this is a great place to be, and I am very excited to be here.

What about your interactions with other Universities?

That is a great thing, because the CIC or the Big Ten has universities that very closely collaborate. In fact, Marianne, you just drove over in an hour and a half from Illinois. We have great collaborations with Illinois, Michigan, of course. We are two hours from Chicago, so it is an opportunity to work with many researchers. I mentioned the ones at Purdue, but the whole region is quite a rich area, and a great place for people to thrive in databases.

What led you to get an MBA in 2002?

When I went to Duke, which was to become department chair, it was just a great experience. It was an experience of building a new department culture, fundamentally based on getting everybody involved from the students on up and energizing it to really move from where it was to the great department it is today. In the process, I got very interested in academic administration, which I think is really computer science on a grander scale. It is problem solving, or to put it more positively, it is finding solutions. And I wanted to get a more formal background. An MBA was really an eye opening experience, because it is a new culture, you are learning new tools, and it was just fascinating to me, especially this notion of strategic planning, which is so important for what we are doing now. So, I just had a great time there. Plus, the Fuqua School at Duke has absolutely the best food in Durham, and we could eat all we wanted, so it was worth it just for that alone.

You mean the MBA students have free food?

Yep, they sure do.

Maybe we should try that in Computer Science.

Well, it might be costly, if you have ever seen the grad student receptions, but I am sure it would be effective.

So, how did you have time to do the MBA while you were also chair of the department?

I timed it so that it was near the point that I was going to step down, so I really overlapped just a semester that way. Then, fortunately, I taught half-time during the following year, so it really worked out well. It was a lot of work, but it was a great experience.

Has your MBA been useful?

Oh, definitely. One thing is just the way that it helps you look at problems and situations and understand the inner relationships, but just thinking strategically and long-term and how you need to really focus on what is going to count down the road because when you get there you cannot go back and change things years ago. We are in the midst of strategic planning now, and one of the things we did that was really fundamental that I think is quite unique across the country is that we have instituted a way of dealing with these large multi-disciplinary problems that are society-wide: trying to find new forms of energy, trying to deal with the climate change and the environment situation, trying to cure and prevent disease. These are problems that require contributions from multiple disciples; certainly computer scientists, but from all over. They just were not getting proper attention, because we were doing things discipline by discipline, and we were focusing on hiring faculty who were going to be the best for our individual disciplines. And in fact, if a faculty wanted to work elsewhere and collaborate, they were almost seen as perhaps a department losing half of a slot, so we wanted to allow departments who had these priorities already to be able to realize them.

We spent a year determining the priorities, but we also had a mechanism in place so that as we were growing — and Purdue was growing by 300 faculty, 60 in our College of Science — and filling these positions, we adopted the approach that we were going to devote these multidisciplinary priorities as the key for these growth positions. We did college-wide searches for these areas, and it's become so much a culture now at our college that as we near our steady state in faculty size, we have decided this is something we want to continue, but we have to do it by a different mechanism. The MBA experiences now help me help design the new mechanism because it is a different circumstance; you cannot use the old approach. You have to design something that makes sense for the time. So we have that, it is unique, it's for our current situation, but it is allowing us to continue this multidisciplinary momentum. So that is what an MBA can help do.

You're now in your sixth year as Dean of Science here at Purdue. What do deans do?

Well, our fundamental mission is to help faculty, students, and staff succeed, so that is my number one goal; and it is through visioning and strategic planning like I talked about. It is raising money. It's trying to be careful in budget management so we can spend money for the things that are important. It's designing curriculum. It is really helping people succeed, fundamentally.

But everything you have just said, at least at Illinois, is also the job of a department head.

That is true, but deans have a broader responsibility. They need to help facilitate the interactions between departments, which is really a substantial challenge. It takes a lot of collaboration and listening. You have really got to communicate and talk a lot with people to understand where they are coming from, what they want to do, and how you can best help them succeed. It is a big job, but it is really fascinating, because when things work, they can have a dramatic effect on people, on lives, on jobs, on revitalizing a state's economy, hopefully leading this country to a brighter future.

You have 5 papers in DBLP for 2007, and more than that for the previous year. How can you be a dean and still be doing research?

So what you are saying is that I am actually publishing less as the years go on, is that what you are saying, Marianne? (Laughing.)

It actually goes up and down, so I don't think we can just extrapolate linearly.

I think, to me, I love research. But more fundamentally, I think it makes me more in tune with what is going on in the college. Staying involved in research keeps me vital. Faculty work incredibly hard, they have a lot of things pulling them in different directions, and I think I should at least work as hard as they do, because we have such a great group here.

People always point to the physicists saying how effective they are at working together to get funding for their research. Computer Scientists tend not to do things too often as a body, or speak with one voice.

In fact, they often shoot each other! I guess that is a way of having one voice: if you shoot each other, there is only one person left. Astrophysicists, for example, are renowned at getting together, deciding what are the key often instrumentation needs that they have that will enable the great things they want to do. Then, in a single voice, they lobby and get those sorts of things. That is really what the CCC is all about. Ed Lazowska is leading that effort in the CRA. It is very important to our future because we need absolutely to get that message out. We need to address the pipeline issue. We are seeing slightly higher enrollments now, but we are 50% under nationally in enrollments in computing than we were just six years ago. It is quite a problem. So we have to get the pipeline in because when you look at the *Gathering Storm* report that came out of the National Academy, there is a tremendous need, and computing has one of the most opportunities for jobs of any discipline. We have 150,000 new jobs created each year, and we graduate 50,000 students.

I think the most important thing is to go and talk with your colleagues in physics, chemistry, biology, history, music, other parts of engineering, because they are just ripe for applications and new kinds of insights that will help motivate new things.

You wouldn't know it to read the newspapers, would you? They always talk about off-shore jobs.

Exactly, I think it's parents telling their kids, "Don't major in computing because the jobs are going overseas." So we are trying to get the message out that it is actually the opposite. And unless we do something, we are going to be struggling in this country, and the biggest place we can make a mark is in the under-represented groups. For women, we are down tremendously for women going into computing these days, and minorities, such as African-American, Hispanics, and Native Americans, we need to do a much better job. And southerners too.

Many people think that computer science as an academic discipline will wither away like railroad engineering: today, you don't see Departments of Railroads in universities. Recently, computer science has been moving closer to its application domains, and you can see this trend especially clearly in the database world. Are we going to wither away and be absorbed by these application areas?

I hope not. And I think the key to being a vital field is to actually embrace those connections and make them a fundamental part of what we do. The real value of multidisciplinary opportunities is, first of all, that they solve the big problems, not artificial problems. Secondly, the most effective outcome is when you really make deep contributions within each discipline as part of this collaboration. And in the course of working on these problems, you will have suggested to you fundamental problems in your discipline, and that is what keeps disciplines alive. If computer science can really embrace this collaborative role it has with other disciplines, it will be revitalized by the very issues that those other disciplines suggest, and that will always keep computer science as a very strong force that will warrant and have people's appreciation.

The way you say that, it almost sounds like the other fields will inspire us by suggesting what direction we should be going, rather than CS having the intellectual leadership.

Well, it is a collaboration, and I think it takes the trust and willingness to not be concerned about who suggested what, so that you can just drive forward, and collaboratively both groups — application arena groups and CS people — are going to make fundamental contributions. If we don't do that, I think what will happen is that the other disciplines will recognize the need for it on their own and adopt computing in their disciplines, and I think that's what the real danger is to computing. So we have an opportunity to revitalize computing by embracing all of these opportunities.

What are the most challenging database issues in other scientific disciplines?

I think in the arena of the life sciences and biology, there are great opportunities that put databases at the fore. For example, in biology, I just have to mention that here at Purdue we have what I think is the top structural biology group in the world. They are focused on understanding the geometry of macromolecules, whether they are viruses or nucleic acids, or whatever, because, in biology, form often determines function. If you take this virus, and you can understand its structure, then drug designers can design drugs that bind to it just right to block its function and cure the disease. Bringing geometry in a fundamental way into databases is really an important challenge — and a very necessary one for this huge area of life sciences. I think that is a great opportunity. Other applications where, for example, satellite data come so fast, suggest new ways of approaching databases, like data streaming, those are interesting aspects too. So I think there are a variety of ways where databases can grow into new areas.

Very few computer science researchers come from the deep south in the US — although you and I are two exceptions. What does your southern background mean to you?

Well, as you know southerners have just in them an identity, and it is especially true in New Orleans because of the very distinct culture that is quite different from the rest of Louisiana, for example. So I will always consider myself a southerner. I am concerned. I think the south has suffered because it is not participating in the high-tech revolution that other parts of the country are really deeply involved in. We need to reverse that. We need to get all under-represented groups involved because we have this great shortage, and this is an opportunity to try to tap into the south and get them focused. So, as a southerner, I feel a lot of regional pride, but also concern, and I hope we can help reverse that situation.

Our fundamental mission [as deans] is to help faculty, students, and staff succeed.

So when you talk about tapping into it, do you mean we should take those southerners and bring them up north and educate them in the ways of computers, or are you talking about a revolution from within?

Certainly at southern universities there are great opportunities to develop a more substantial database presence, and in general computer science. That, it think, will be very important. As they develop new technologies, they are going to need that environment. Richard Florida is an author who has this thesis that the great economic centers are fundamentally built around great universities because creative people are attracted to places that are vital in culture. We have to build that in the south, and I think it will all come together.

Southerners are attracted to places with great football, so maybe that is key.

That's true. I went to Notre Dame as an undergrad, which is an archenemy of most schools in the south, but it was a fun rivalry.

So, if you have this strong southern identity, where is your strong southern accent?

Well, I have no doubt lost some of it. The best book to get an understanding of real New Orleans is *A Confederacy of Dunces*, by John Kennedy Toole. In the forward of this book, there is a little blurb from probably a hundred years ago that describes a

New Orleans accent as really a soft Brooklyn accent. And that is really what it is. If you go to New Orleans, if you hear a southern accent, it is certainly someone who wasn't born there. But a real New Orleans accent is a real Brooklyn type accent.

Can we get a demonstration here? I'm not quite following you.

Well, if I saw you at the local drug store (and of course you'd have your hair up in curlers), I'd say (in New Orleans accent), "Hey, where y'at, MariANNE? Whatcha doin'? You wanna go get some red beans and rice?"

There is that tang in there.

But it is nothing like a southern accent. In fact, an expression in New Orleans for "how are you?" is "where y'at?" New Orleanians are called Yats as a result. That's the name of the restaurant that just opened here in West Lafayette; the New Orleans restaurant is called Yats.

Do you have any words of advice for fledgling or midcareer database researchers or practitioners?

I think the most important thing is to go and talk with your colleagues in physics, chemistry, biology, history, music, other parts of engineering, because they are just ripe for applications and new kinds of insights that will help motivate new things.

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

Actually, it would be to go home and spend more time with my family and kids. I have an incredible wife, Sharon, and three wonderful kids, Jillian, Scott, and Audrey. I just wish I could say I was more responsible than I am for how they have turned out. So I would spend more time at home.

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

I just wish I had the time to learn more things, because there are so many fascinating connections, and many things that I do are dealing with applying paradigms or insights that I picked up one place that shed a new light in another domain and lead to interesting new results. I just wish I had the opportunity to learn more things and keep up with all the things going on in computing and other fields.

Well, thank you very much for talking with me today.

Great, it was a pleasure to be with you. Thank you.