Christopher Ré Speaks Out
on His ACM SIGMOD Jim Gray Dissertation Award, What He Wishes He Had Known As a Graduate Student, and More

by Marianne Winslett

Christopher Ré
http://pages.cs.wisc.edu/~chrisre/

Dear readers, you may have noticed that the printed versions of these interviews are not appearing in every issue now. I took a position as the director of UIUC's new research center in Singapore (http://adsc.illinois.edu), and that has reduced the amount of time I have available for other activities. Yet the video versions of the interviews still pop up on the SIGMOD web site regularly – so what is the problem?

The difficulty is that spoken English is almost a different language from written English, so the conversion of an interview transcript into a coherent document is a labor-intensive process. I am looking for a second editor who can take over this process, similar to the video editor who produces the video versions.

The new editor(s) of the written versions needs to be really good with English, because the written version must make it perfectly clear what the interviewee meant, while at the same time not changing the interviewee's unique verbal style. And there are delicate judgment calls to make, e.g., if the interviewee sounds extremely Italian, that should still show up in some way in the written version. To understand the task, try listening to a video version of an interview while reading the printed version. The reward for taking on this new position is to have your name in the byline above. So, if you are interested, please send email to winslett@illinois.edu !

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 we are at the SIGMOD 2010 conference in Indianapolis. I have here with me Chris Ré, who is an assistant professor of computer science at the University of Wisconsin - Madison. Chris is the recipient of the 2010 ACM SIGMOD Jim Gray Dissertation Award, for his dissertation entitled Managing Large-Scale Probabilistic Databases. His PhD is from the University of Washington. So, Chris, welcome!

Our runners-up for the award this year were Soumyadeb Mitra (University of Illinois at Urbana-Champaign) and Fabian Suchanek (Max Planck Institute for Informatics).

So, Chris, what is the thesis of your thesis?

The thesis of my thesis is that it is possible to build large-scale probabilistic databases. The reason you would want to build such a crazy thing is that there are a lot of applications out there that are managing data that is increasingly large and increasingly imprecise.

Can you give me some examples?

Certainly. One example is RFID data, where you are trying to track objects and people as they move through space. Whenever you go to measure the physical world, you have issues like measurement error, so the resulting data is much less precise than the data we are used to having in traditional relational databases. And there are many other examples, such as information extraction, or data integration, where the data is somehow just less precise than what we are used to putting inside a database.

What are the research challenges in managing this type of data?

There are many challenges, and I was fortunate [to win this prize], because there are a lot of people working in this area. The main challenges I focused on were scalability and performance. In general, the way we model all these imprecise information sources is to use probability theory. Probability theory is great, because it lets you talk about many different kinds of imprecision in one unified way. But then, when you go to actually process a query, you have to consider all these different alternatives. And when you consider all these different alternatives, you are doing a lot more work than you would do in a traditional relational SQL style query processor. So I focused on performance and scale, and that is really the contribution of my dissertation.

It sounds rather impossible! What is the secret to be able to manage all that?

The secret sauce, I think, is a lot of other smart people's work. We took techniques from all over the place: we borrowed techniques from AI literature, machine learning literature, logic, probability. I had great co-authors around me, and we mixed all that stuff together. So some of the contributions that were in my dissertation were just observations, such as that in a lot of these applications, people are only interested in, say, the top five or ten answers. So when we are building a system that can handle imprecise data, we don't have to compute very many answers to each query. Once you have the idea that people want to focus on just the most likely answers, then you can zoom right in on that problem. We had some query processing techniques that allowed us to address that problem. The other thing we had in our bag of tricks was classical database tools, like materialized views, and some ideas from approximate query processing, popular old topics that we could bring to bear on probabilistic processing. So, the secret sauce was really other people's work. Ha!

Okay, so you claim! Although, since you got the award, I'm sure there are a lot of new results in your dissertation. In fact, the AI community and the logic and probability communities are not known for their attention to scalability.

That is one thing that I was really excited about in my work. I was applying almost a traditional database focus to these classical problems from these other fields. It was really a joy to pursue, and I hope that some of the people in those fields are excited by this work too.

What do you know now that you wish you had known as a graduate student?

The one thing I wish I had known was how much of the learning during the PhD was by observation. I had a great model when I was doing my PhD: Dan Suciu, my advisor. Just watching how he did things, and imitating him, was a great part of the process. But now that I have gone through the PhD process, I see that I didn't realize how many things I should imitate from the other people inside the field. Everything from how they communicate, to how they write their papers, to how they conduct their experiments, down to really fine details. There is really no way you can tell someone all those things, they just have to pick them up. I didn't realize how much of that was on the graduate student agenda until I got into my second or third year, and then all of a sudden I understood.

How did you get into the database field?

That is an interesting story. As an undergrad, I was a math and CS major. And then, on a lark with a friend, we decided to take the database course at Cornell, which is CS432. Jay Shanmugasundaram and Johannes Gehrke were team-teaching it that semester. And they had a guest lecturer, Jim Gray, and he was talking about the World Wide Telescope project. I remember being blown away by the simple answers he was giving to complex problems. As a mathematically minded undergrad, this was just amazing to me. I really fell in love with the topic as a result of that talk, so to me it is a real honor that his name is on this award.

I have a kind of similar story, only the teacher was Phil Bernstein. I was also a math major, and Phil made databases seem so interesting to me as an undergrad.

It just somehow came alive. They had the simple answers to complex problems that just blew me away.

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

Probably I would be a bit more disciplined. I tend to get very interested in a lot of different topics, which has been nice in some ways. For example, I have a thesis that touches a couple of different areas that I really like, and I am proud of that thesis. But I get distracted by topics pretty easily. There are a lot of fascinating topics in data management and if I could stick and hold on to one, I would be pretty happy about that.

Thank you for talking with us today!

Thank you so much!