Rick Hull Speaks Out on Asking the New Question

Marianne Winslett and Vanessa Braganholo

Rick Hull

http://researcher.watson.ibm.com/researcher/view.php?person=us-hull

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 Snowbird, Utah, USA, site of the 2014 SIGMOD and PODS conference. I have here with me Rick Hull, who is a researcher at IBM. Before that, he was a professor at the University of Southern California for many years. He also managed a research group at Bell Labs, where he was a Bell Labs Fellow. Rick is an ACM Fellow and a coauthor of the classic database theory book Foundations of Databases. His Ph.D. is from Berkeley. So, Rick, welcome!

Thank you!

What could database theory researchers do to increase their impact on the world?

Well, that's a big question to get started with. I think the challenges of data management today are really expanding from what they were 20 and 30 years ago. Data is so pervasive in our lives today, from social media to ecommerce, and things like using weather data for smarter farming, it goes on and on. We've been discussing this in the PODS community broadly – how to study the fundamental issues raised by new kinds of data and new uses of it. I think the main thing is to consider what are the problems today and then what are the techniques that could be brought to bear. So rather than using mathematical logic as the starting point for most explorations, it is time to more fully embrace additional frameworks, including probability, statistics, etc.

But people have been telling me that there is still plenty of room for logic.

Oh, absolutely. I'm not saying that the logic is no longer needed. I am just saying that we need to expand the full range of techniques that we can bring to bear, the kinds of models that we might look for. The logic is still a very important foundational element. I know, for example, some of the basic techniques in the statistical approach start with counting all of the true first order models of a given theory.

[…] with that paper, we were asking a new question. This was part of Seymour Ginsburg's mantra: always ask the new question. It was an unusual question, a non-standard question.

And is that happening now? Already happened? Or is that the next step?

I think there are very positive signs in the past 2-3 years. Just at this conference, we had the Big Uncertain Data workshop, which was a very deliberate attempt to bring together people from the database theory side (especially work on probabilistic databases) with people from the machine learning side.

What about the fact that there aren't that many commercial probabilistic databases?

It's funny you should ask that. I am also concerned that in the old days the data management activities, including those that the database theory community was contributing to, was a growing industry. It was really a growing, very important field. Now, the big growth seems to be in the big data, the analytics areas. The machine learning community has been producing artifacts that are useful to that industry, and I'm hoping that over time the database community broadly and the database theory community can contribute into that area as well. There is also important work in other areas of data management, for example multimodal data, incomplete data, data-centric workflow.

You stayed at Bell Labs for a very long time. What did you enjoy about your time there?

It was about 12 years. Especially at the beginning it was a very exciting place. The company was growing. It was owned by Lucent at the time. It was my first experience in an industrial research lab. It allowed me to do both -- to continue with theoretical research and advanced research, but it also gave me a chance to be talking with customers, wrestling with the challenges of how you take ideas and bring them into reality. How do you bring some kind of value or capability or something that will be used by the average person on the street?

So that would be true of all industrial labs?

Each lab is different. You know, I'm speaking of my experiences with Bell Labs and then IBM Research. Both I think offer this breadth of opportunity.

And you haven't gotten back to academia, so it seems like you're voting with your feet that you really like that connection to the customer.

You never know what might happen. I think the opportunity to work in a larger group as well as continuing with individual contributions is something that I enjoy.

About the Alice book, that's the nickname for your database theory book. What was it like to write that book?

That was a lot of fun. To understand the context, in the early 1980's I was at the University of Southern California, and Victor Vianu and Serge Abiteboul were there as well, and Seymour Ginsburg was the mentor for all three of us. This was right at the beginning of database theory and so we felt like we were on the ground floor. We were learning some of the early results. We were under Seymour's guidance, starting to build up our own body of results. So we wrote the book about ten years after being together at the University of Southern California.

When we were writing the book, it was a point where the foundations of database theory, at least that the first real era of database theory was, I feel, coming to a kind of closure. Well, not a closure, but there was a feeling of completeness to what had been studied. So the book was at a perfect time to capture and encapsulate that body of work and hopefully provide the foundation for the next generations of work.

Well, that's a good point because a reviewer on Amazon says that although it was published in 1995, it quote "it is still the gold standard… especially in consideration of the fact that nothing much has changed in database technology in the past 30 years or so". Do you agree with that?

No, I wouldn't agree. I mean it's nice to think that it's a gold standard for something. Maybe it is potentially a gold standard for that period of the database theory and the basic logical framework that was set up. At the same time, since then there's been a tremendous body of work in database theory to understand XML as a major area, connections with XML, automata, constraints, etc., further advances in constraint databases, description logics and data, and of course now as we go into the big data period. So there has been a lot of advancements.

Students often choose one of your papers from the class reading list because it's shorter than the other options and then they just knock themselves out trying to understand the paper. For example, many researchers have been influenced by your PODS 1984 paper about when two databases are equivalent[1]*. What's so hard about that topic?*

(Laughing) The information capacity paper and its follow-ons… Yeah, with that paper, of course, we were asking a new question. This was part of Seymour Ginsburg's mantra: always ask the new question. It was an unusual question, a non-standard question. I think that's why conceptually it has been hard for people to think along that line. Secondly, we had four levels of relative information capacity and each level called for some different techniques. So I think that maybe that also makes it harder than some other papers where there's kind of one core technique and then it's just played out.

I think that whole direction was really important. Nowadays, when query answers have some sort of statistical aspect to them, in my group, we believe that if two databases are equivalent you should get the same answer no matter which one you run your algorithm over, which is kind of a radical notion. But we really believe that should be true and if you can't talk about what's equivalent you can't argue that you should be getting the same answer. So that's a nice example paper I think to pick out. It was back in '84 but it's still important 30 years later.

What are artifact-centric business process models?

That's a big question. Maybe the last four or five years of my research work was in that area of what we call business artifacts, or using business artifacts to support business process models. So business artifacts are really a great opportunity for the database community and others to study a combination of data and process. Kind of married as equal partners. You see traditional business model management is focused on the process side; flowcharts or maybe it's based on Petri nets. A lot of the research in that area has focused just on the process and it has left the data as a second-class citizen. But in reality, the business process is really touching data right and left. So with business artifacts, the core model really focuses on what we call key business-relevant conceptual entities. Key conceptual entities that progress through a business during normal activity. So as an example, we talk about the FedEx package delivery. Not the package, but the package delivery, and think of it in terms of when the package was first received by FedEx, the transportation, the delivery, the sign-off, also the billing, how that goes. With a business artifact model for that, you have an

[1] Richard Hull: Relative Information Capacity of Simple Relational Database Schemata. PODS 1984: 97-109.

information model that holds the relevant data that may be obtained as that entity progresses through its process. Also, you track the life cycle model, the possible ways or possible activities that might happen to the package.

I would argue that the database community does the reverse of what you're saying the business process people do. We care enormously about the data and we don't care at all about the process. So, is this where we're supposed to meet, in the middle?

Definitely, it is one opportunity for the two sides to meet. What we found is that the business artifact perspective gives a very strong intuitively natural top-down view of the business processes. Typically there are 3-7 business artifact types that you need to model a given process. They can be cross-cutting. So if your business process is cutting across multiple different silos of your business, often the business artifacts span multiple silos and give that top-down end-to-end view that's lacking in so many other cases. Let me say that there has been a body of research. There are probably now 20 or 30 active researchers in the database community, the AI community, and the business process management community that have been working on this model and its marriage in areas from efficiency and distributed systems all the way up to verification, really spanning the gamut from systems to theory. Actually, the theory side was discussed in the Diego Calvanese's PODS keynote talk last year (2013).

Is it getting traction in the business world?

I would say absolutely. The work on business artifacts, which started at IBM Research in 2003, was, in fact, the motivation for me to go to IBM Research. By about 2011, people came to realize that business artifacts were actually very much a formalization of the case management approach to business process modeling and now the work we did at IBM Research has provided the foundation for the OMG standard on case management and also for the IBM case management product.

Good! Great, it's great to see that happening. Speaking of IBM, what is next for IBM? They've stayed alive so long so there must be something new just around the corner.

Well, they are very deliberate at the corporate management level of steering the ship, always thinking about what is next. You know we saw recently in the past several years this initiative around this smarter planet, smarter cities, smarter education, and smarter healthcare. Recently, the big topic area both in terms of activity and in terms of the marketing is on cognitive computing, as we call it. It's building on the success of the Watson deep question and answer system. People may recall it had been featured on the Jeopardy television game series maybe a couple of years ago. There, it played against two Jeopardy champions and it demonstrated the ability of a machine to have processed just tons of both structured data and unstructured data, to be able to reason about it, and to be able to, in this case, formulate questions based on all of that learning. Now there is a division of IBM that is focused on Watson and applications of the Watson technology. The research division has also been reorganized a bit and there's now quite a large activity around what is the future of computing given that it can take advantage of this unprecedented amount of processing power and in particular, processing of the unstructured data.

I enjoy being with people and being able to have a diverse set of challenges in front of me.

What kind of applications are we likely to see coming out of that?

I think we're already seeing some of them and they'll just get stronger. I mean one area that even before it was being labeled cognitive computing is in the smarter healthcare area. For example, they are training the Watson system to be able to take the medical board examination. After medical school, the doctors take some kind of exam. Well, they're training Watson to be able to take that exam and also to explain the reasons behind whatever answers they are giving. There's also an activity where IBM is partnering with Sloan Kettering, the Cancer Care Center, to help with cancer diagnosis. That's one area.

Another area that has kind of personally been intriguing for me is in smarter education: enabling students to experience personalized learning pathways. This means they can be working on material that is delivered over a tablet and through the use of analytics, through deep analysis of text material, of the problems, etc., you can really deliver to the student the next best module for that student, his learning style, what he

knows, where he is trying to go, in terms of his academics or his career. We're also seeing it in more business settings, financial analysis, for example, advising people on where to invest their money, how to build their investment portfolio as they move through their lives.

That could really be beneficial for math in the K-12 era education where kids think it has no application to whatever they're interested in and that is so completely false. So if they're interested in construction, there is tons of math in construction, if they were interested in baking or sewing, there's tons of math in that and if the problems they were given whether it's trigonometry, algebra or whatever were tailored to their interest… same formulas but expressed differently then they would see how it connects to the real world, but we don't do a good job of that. Or if we do it's about trains traveling in different speeds and different directions and where they will collide or whatever.

> ***[…] take the time to really work a problem, think about the problem, try to go deeper, try to ask the next provocative question.***

Those are really good examples actually because the idea is that you can start to tailor many aspects of what is being taught to the interest of the kid and also to their aspirations.

Okay, let's see. Have you found it more satisfying to do research or to manage research?

You know, I think it's really the mix that is most exciting for me. I like to have my hands into something concrete, even if it's a mathematical abstraction, it is in a way concrete for me and you're working puzzles with it or you're trying to figure out an algorithm that's going to work or prove that something is correct. At the same time, I enjoy being with people and being able to have a diverse set of challenges in front of me. So in some of my most enjoyable periods both at Bell Labs and IBM Research, that's been the experience: I've been managing, and I've been collaborating with outside universities, maybe working on a project with a customer, but also working out some little theorem to help solidify the foundations of the concept.

Are you still collaborating with Serge and Victor?

Not as much as I had. With Serge, it's been a while. With Victor, he has been involved with the business artifact work. In fact, with Victor and Alin Deutsch, we've started a line on verification of business artifact properties and so that's been quite enjoyable.

So the take home message for all the students reading is that whoever your colleagues are on your grad school days you may still be working with them many years later, so those relationships can really last a long time.

You are perhaps the coolest person in the database research community. Where did you get your cool?

That is a funny question. I wonder where you get those questions… I'm glad at least you think I might be cool. You know, I'll give you three possible factors. So, one is when I was young, my father was into camping, the outdoors and we would go camping or canoeing on the river and be outdoors for two or three days at a time and I think that kind of experience of being in nature (this was long before cell phones, but it was also a time away from a lot of distractions) it's kind of an interesting mindset to carry with me. Of course, being an undergraduate at the University of California Santa Barbara, on the beach, that was a big one. A third one is, Europeans have a certain cool and coming back to Serge and Victor I spent a lot of time with them and then I was able to visit Serge in France for several summers working at Inria and Victor was typically there. It may be the exposure to Serge and Victor that really put it over the top.

A shared cool. I did not make up that question myself. I got it from one of your colleagues and same holds for the next question, which is: tell us about your hair. Can you show us your hair?

(Rick shows his hair.)

Look at all that hair! Ok, so what's the story behind that?

Well, I haven't thought about that recently. I guess one answer is that in research we're always thinking about something new, the new question, and the new technique. At the same time, I still have my roots. I like the old as well and I grew my hair out in the late 60s. It was kind of part of the peace movement back then, and somehow I never cut it off.

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

I think one word would be networking, get to know your fellow researchers better, try to collaborate, that's just such a wealth of stimulation. I think another thing is the power of the human mind. What I found at least is that if I really live with a problem, work with a problem, make time to think deeply, challenge myself, that's when the mind can really go to the deeper insights. So I would recommend: take the time to really work a problem, think about the problem, try to go deeper, try to ask the next provocative question. I think it's so easy in this day and age to get distracted by the next email, the next phone call, the next meeting or whatever. So, you know, trust your mind and dig deep.

They're under so much pressure to get that next paper out so that they can get their first job or whatever that it can be hard to do.

Yes, I agree, and I remember actually that the paper on relative information capacity was published in PODS then it was time to write the full journal version, then there was a deadline, and I realized there was a bug (you know, a minor bug). So I spent some long nights wrestling that to the ground and on the one hand that is fine, I did have the chance to think about it deeply and met the deadline, but it was unfortunate to feel under that pressure.

That means that if they choose that PODS paper for their class because it's shorter, that they should if they really understand it, they should find a bug in there.

I wouldn't want to go on record saying that they should find a bug there, but I think a challenging exercise may be for a very motivated student would be, "What's the difference between the journal version and the conference version?"

Ok, very good!

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

Well, in addition to doing some more of that deep thinking that I don't seem to get a chance for anymore, I would say reading other people's work, I find that I just don't have enough time to read other stuff and to work with other people's material and really have a strong understanding of it.

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

In my case, I grew up through a math major and my Ph.D. was in math, although formal language theory, so from a very mathematical perspective. As time went on, I became more involved with a system side as well as the theory side. I think a change would have been to spend more time on the real computer science side of things, programming, programming languages, abstraction. These are principles and foundations of our field that it would be nice if I understood them a bit better.

Among all your past research, do you have a favorite piece of work?

Well, I think it's the artifact-centric work broadly, but if you wanted me to just pick out one paper, I think what I would pick out is the paper we had in the ICSOC 2009 conference[2]. That's the International Conference on Service Oriented Computing, and the paper is on artifact-centric hubs.

Thank you very much for talking to me today, Rick

Thank you!

[2] Richard Hull, Nanjangud C. Narendra, Anil Nigam: Facilitating Workflow Interoperation Using Artifact-Centric Hubs. ICSOC/ServiceWave 2009: 1-18.