H V Jagadish Speaks Out on PVLDB, CoRR and Data-driven Research

Marianne Winslett and Vanessa Braganholo



H V Jagadish
http://web.eecs.umich.edu/~jag/

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, cite of the 2012 SIGMOD and PODS conference. I have here with me H. V. Jagadish, who is the Bernard A. Galler Professor of Electrical Engineering and Computer Science at the University of Michigan. Jag has served as the editor-in-chief of the Proceedings of the VLDB, the database area editor for CoRR, and a board member for the Computing Research Association. Jag's PhD is from Stanford University and he's an ACM Fellow. So, welcome Jag!

Thank you, Marianne.

Jag, you've been very involved with the Proceedings of the VLDB. How did that get started?

The Proceedings of the VLDB just happened to come together. This is really how it happened. I was part of the VLDB Endowment Board of Trustees and there had been a lot of discussion amongst various people who had been on the board before me about publication models and what one should do. There had been a broadening effort that Phil Bernstein and others had been pushing. There were some who felt that conference publications did not get the same level of respect as journal publications, particularly in some countries. There were others who were concerned about how we did our reviews and what our reviewing processes were. Somehow all the vectors lined up at the right time and I just happened to be able to make use of all the forces that were there at that time. When it suddenly happened, it was actually very quick. There were a number of pieces that needed to come together for PVLDB to happen and all of that happened within one VLDB Endowment board meeting, which is typically not the way that VLDB operates. That's because there had been many years of preparatory work and so people sort of knew all the issues. There had not been partial solutions before, they were just issues, what we should do, and inconclusive discussions. When there was something that seemed like it had a chance of working, I think the trustees were very enthusiastic about trying out the experiment and seeing how it worked.

[…] a great deal of the work that people are doing in a data-driven manner in many disciplines is often prey to all kinds of biases and errors. It's very easy not to have enough statistical power.

Is this experiment helping with the communities that want to see a journal publication? So is PVLDB a journal?

PVLDB is a journal when its advantageous to be one. I think that it is truly a hybrid. It is not a standard conference publication. It is not a standard journal publication. For those who keep books, the fact that we have an explicit ISSN, which is what makes it a journal as opposed to just an ISBN, which is a one-shot thing that each conference proceedings gets, makes it technically a journal for classification purposes. Other than that, I think that the nature of the publication, and the spirit of it and the way in which one reviews it and evaluates it, is very much in what we think of as conference style as opposed to journal style.

And is it in that ISI index that some countries rely on?

The ISI index is one place for instance where having an ISSN is very important.

How is the impact factor looking for it so far?

It's too early. It turns out that Thomson Reuters has a process for putting things into their index and among the things they want to see is a minimum bar of three years of publication history on schedule**.** So apparently they deal with a lot of publications that do not end up having enough volume and so the issues get delayed. Now, we are all used to backlogs in our journals and people are trying to minimize the backlogs and editors are wheedling the extra pages from the publishers, and things like this. There are places where journals just do not know how to fill their pages. They are promised a quarterly journal but they have a hard time actually bringing one out. So for whatever reasons, that is a rule and we have not had three years of regular publications[1]. So we are not yet indexed in the ISI, for instance.

So you've been very involved with this CoRR (Computing Research Repository). What's that all about?

The Computing Research Repository is something that some people put all of their work in and others may have never even heard of. It seems to have an uneven uptake in our community. The idea behind this is that there is a central place where people put any work that they think others may want to read. There isn't a review process other than for appropriateness. So we do talk about the category and we want to make sure that if you claim that your paper is about databases then it is about databases. So it isn't just that we keep a political creed out, but one of my jobs as the database section editor for CoRR is to make sure that a paper that is about "software engineering" doesn't have a

[1] This interview was conducted in 2012.

"database" label by mistake. Sometimes there are questions about where exactly do we draw a boundary, does a paper deserves two labels, and things of this nature. Anyway, the thing with this is simply to have one place where people can find work on a particular topic. This is something that physicists have had great success in using extensively. I think most areas in physics do this and some areas in mathematics and other fields. Computer science was not even the first to the game here, but the same infrastructure is being used for computing. Some people and some sub communities seem to have adopted it with gusto and others have just totally ignored it.

I think that the value, if there is good adoption, is that it saves you the effort of doing things like a web search to find papers. We have organized collections for things that have been published in good places (things like our SIGMOD DiSC -- Digital Symposium Collection) and things of that nature, which for our sub community works very well. So the need for the database community for something like CoRR may be less. We note though that CoRR isn't just for things that are published in good places. It is for everything.

I think of it as a pre-print place or a pre-submission place.

Well, so one of the things we've also done is when papers appear in PVLDB, they also get deposited in CoRR. So it is not something that gets there as a pre-print, it is put there at acceptance. And many conferences, workshops and journals do this with CoRR on a regular basis. When papers are accepted, there is a batch upload.

The people I know who are enthusiastically depositing and looking in CoRR are using it as a place to put in their papers that usually they haven't even submitted yet so they're sort of at the tech-report stage. That is different from what you were talking about a minute ago and that usage in my mind conflicts with the double-blind reviewing philosophy. Do you have any comments on that?

Yes, I agree with you that there are people who put things into CoRR at a very early stage, at a pre-print stage or to establish "first in time" for some idea. But I view this as not much different from people putting out papers on the web. There are a lot of people who put up tech reports on the web and I think that double-blind reviewing is impacted even if there weren't CoRR, just because there is a web and there is web search. That is a challenge for double-blind reviewing. I don't think that CoRR makes it that much worse. I personally believe that changing how we manage our archival to support the blindness of reviewing is the tail wagging the dog. I think that whatever one might want to do with respect to whether we use double-blind reviews or not, that is a concern with respect to how we evaluate papers and that should be a second order concern with respect to how we disseminate knowledge: how we store and share knowledge should be the primary concern. Even if it were the case that it mattered that there was a negative impact on double-blind reviewing, I would still say CoRR is dealing with a more important issue than double-blindness does.

So given that we have this IEEE, ACM Digital Library access, do we also need CoRR?

I think that the digital libraries are very good and actually in terms of stuff that I really use in my research there is very little that isn't in either IEEE or ACM's Digital Library. Again, there are some issues: they are both proprietary (they are owned by professional societies), they're available for a fee (so they're not free to use), and they have standards that they have in terms of what material is included. And so things that don't make the cut with respect to the venues that get incorporated wouldn't be in there, whereas CoRR just has everything in it. The other point is that, to the extent that people put pre-prints or tech reports you get faster dissemination. For conferences, for example, things get into the digital library usually several months after the conference, it isn't even at the time of the conference.

One side effect of all the work we've been doing in the database community over the last 50 years (according to Rick Snodgrass, we're 50 years old now) is that scientists have huge amounts of data available to them that they didn't have in the past. How has this changed the way that they do science?

The way I think about this is that the standard way of doing science is what is called hypothesis-driven. You first pose a research question that you're going to ask, you have a hypothesis and then you do one or more of the things that you just mentioned -- let's say an experiment. The result of that experiment will either verify the hypothesis or refute it. And that's the classical scientific method of doing research. The thing that has now become possible is not to have a hypothesis but to have a goal that says "I'll find out something of interest in this space." So if one were to take a cartoon picture of data mining, as we would have talked about it even 15-20 years ago, we would simply say "Well, we have a lot of data and we look for patterns in the data." Let's say we stop there. That is what one can think of as data-driven scientific

research. One can say "I am looking for genes that have some role in some disease. I don't have a clue about anything, except I know how to do sequencing and so I'm going to take a bunch of people who have this disease and a bunch of people who don't. I don't have a hypothesis other than to say that there must be some genes that are different. Then I'm just going to run their DNA and look at where the differences are." This is not hypothesis-driven research. You can state it in terms of a hypothesis, but it is not a very interesting hypothesis.

In the example that I just gave, there actually is a data generation face to the research. One could do this with secondary data. One could say "I'm going to make use of other people's data that's published and do a secondary study with that". The point is that we're learning new things without knowing beforehand what we're going to learn. I think that this is very powerful because it decreases the burden on us to specify a hypothesis in advance. On the other hand, if one doesn't have a good explanation, at least in a post-hoc manner, one ends up with things that are intellectually dissatisfying and possibly even statistical flukes. I think that there is need for people who undertake this kind of scientific investigation to think harder from first principles about the statistics and what the likelihood is that they are seeing results that are not the result of over-fitting or the result of just multiple hypothesis testing or some other issue of this nature. I think that the standards of statistical evidence need to be much higher as a threshold for acceptance when one is doing it without a hypothesis.

One problem I see in the non-hypothesis-driven approach is that I'm not sure how well it's accepted by other people in science. So here's a direct quote from someone who is in the medical industry: "Oh those epidemiologists, they just want to go on fishing expeditions".

I think such statements are actually warranted in many situations because a great deal of the work that people are doing in a data-driven manner in many disciplines is often prey to all kinds of biases and errors. It's very easy not to have enough statistical power. It's very easy to have results that are incorrect because of multiple hypothesis testing or because of over-fitting or because of some other bias in terms of the way things were done. There are well-documented cases, for instance, of people showing things like moving objects in the distance through thought. You know, things of this nature which one shouldn't have a scientific basis to expect. Every now and then there is some such paper that gets published. If one conducts enough experiments there would be some case where just in terms of random association, things will turn out the way that you would like them to be. So, when one is considering some small data sample, and saying "Well in a data-driven manner I see this result, therefore it is", I think one has to take that with a very big pinch of salt.

I don't think that the database community is actually doing very much for scientists.

That having been said, I think that there is a question of the comfort zone for somebody who has been trained in a certain way of doing work. As a person who begins with the data, which is what I'm certainly trained to do, I often have discussion with people who are used to thinking about the hypothesis first and just feel uncomfortable at a gut level because they are being forced to think about things in ways that they're not used to. That will instinctively make them react negatively and then it's a question of them thinking it through, with their knowledge, training and wisdom and coming to a conclusion about whether some new piece of work done in a new style makes sense or not.

How well is the database community doing at supporting the needs of scientists?

I don't think that the database community is actually doing very much for scientists. I think that many scientists have a lot of data. I think they struggle with the data and they do all kinds of things with the data that may seem ridiculous to people who attend SIGMOD for instance, but they do it because that's what they know how to do. I think being able to provide tools to support their work, particularly as the amount of data that scientists are dealing with increases, is something we as the community should embrace and I know that at least some segments of our community are thinking hard about things like this. I think we have a long way to go.

Do you have a list of top challenges that we should be working on for the sake of scientists?

Actually, my view is that what we do for scientists is probably not that much different from what we would do for an end user in the consumer arena. I work with scientists as you've said, and I think about things that we might need to do in terms of data management to

help scientists do what they need to do. But when I write a paper in the database world, describing some result, it's usually not hard for me to take the same thing and cast it in terms of a hotel reservation or managing an address book, or something of this nature -- just very simple, personal tasks that end users would do for themselves. So I really think that the challenges are what one would expect if we just sat down and said "Who is using this stuff? What do they need to do?". I think that we get too wrapped up in dealing with what needs to be done inside the box with tightly defined boundaries and I think taking that one extra step of seeing what is it that someone is trying to accomplish with whatever is running on this box would make a world of a difference.

So if I am understanding things correctly, you're saying that the core guts of what they need is already there but it's not friendly enough, accessible enough, missing some layer on top perhaps for them to actually make use of it. Or maybe they don't know it exists…

Yes, to all of the above.

What is the right way to design a usable data management system?

My soapbox position on this has been that usability isn't skin deep. Which is to say, that you can't build a database system first and then throw a pretty interface on top of it and say that you now have a usable database system. Instead, I think you need to start at the beginning from what task the user is trying to accomplish and what knowledge the user brings to the task when they're trying to accomplish this and then see what the workflow should be to maximize their ability to accomplish that task directly and quickly.

To some extent the interface matters, but I think that even beyond the interface, as one thinks it through in terms of breaking a task into subtasks, and what is actually being done, one ends up with an interaction model. This is in effect the query model, the thing we should then have efficient support for. We had better design our database to be able to support that kind of interaction model, not to make people think about it that way. Usually we start with "I got this box and what can this box do?". And so we naturally end up with things that are not particularly usable.

So if I understand correctly then you're saying is we might need to redesign the core, the guts of the system once we figure out what these people really need.

I think that we will need to redesign significant aspects of it. We may not need to re-do all of it. If we get down to things like the actual data store which is, for most purposes, probably not something that would become visible, even there, I can give you a counterexample. So a paper that one of my students had last year[2] was on the system called CRIUS for organically grown database systems where the idea is that the user doesn't have a schema in mind before they start throwing data into the database and so as they come up with new instances, they realize that the schema needed to be richer than what they previously had. So, you start with a single column in a single table, and you grow it from there. Well, even though a lot of our contribution there had to do with how the user does this and what support the user gets and what dependencies mean and how do you keep the user from making errors, etc., the fact that the schema is evolving (and you expect the schema to evolve) on a continuous basis has implications on the kind of storage that you do. So for instance even if you didn't otherwise have a reason to do a vertical storage, the fact that you need to support something like schema evolution might tip the balance. So there are things like this that could affect decisions even at the gut level.

Is database research turning into informatics research?

So a thing I've been trying to do with very little success, when people ask me what area I work in, is to say "I work in information management". Quite often, I get a blank stare. They say "Oh, you mean databases", and that means something to them. I think the reason that I want to say information management and not databases is because, to me, a database is a very specific engine that does something we all understand, whereas information management is the broader universe. Databases have a significant role to play in information management, but I want to lay claim to the broader turf and somehow that has been difficult.

XML query optimization: should we give up and walk away, like we did for relational query optimization and call it done?

I think that at some point things matured to a point that the academic community has done pretty much what could be done. It doesn't mean that everybody should

[2] H. V. Jagadish, Arnab Nandi, Li Qian: Organic Databases. DNIS 2011: 49-63. There is also a more recent paper on the subject: H. V. Jagadish, Li Qian, Arnab Nandi: Organic databases. IJCSE 11(3): 270-283 (2015).

walk away but I think that the bulk of the interest moves on and every now and then there will be some willing person that comes in and changes the paradigm and makes us all think something new about something and such things may happen.

Why should we care about the Computing Research Association?

The Computing Research Association actually is something that does a great deal of good. It is an organization that not too many of us may have that much familiarity with. It is a professional society, except that the members of the society are computing research departments as opposed to individuals as say in the case of something like ACM. What the CRA does is think about what is good to support the computing research enterprise with a little bit of an administrative view much more so than say, something like the ACM. So in terms what they specifically do, there are things that are bread and butter everyday things. They do what is known as the Taulbee Survey of salaries and placements of graduates and things of this nature and this is something that helps us keep track of where things are in the field. Helps us keep track of the health of the field. Helps our department heads fight for larger raises…

[…] usability isn't skin deep. […] you can't build a database system first and then throw a pretty interface on top of it and say that you now have a usable database system.

(laughs)

Yes it does! That is one reason to pay attention to the CRA. But I think beyond this, the CRA is a good place because of the way it is set up to take action on items that are of broad interest to the computing community. For example, the CRA was responsible for a postdoctoral fellowship program that was put into place exactly when the downturn hit about three years ago and jobs dried up. This program is now being phased out as the economy recovers and as hiring is coming back up to normal levels. I think that if the organization weren't there, this wouldn't have happened. I forgot to mention that the CRA is very much North American, so it's not a worldwide thing unlike say, the ACM. So one of the things the CRA does spend significant effort on is in educating government officials on the benefits of funding computer science research. Again, in that, there have been many activities that the CRA has undertaken and the fact that there is generally bipartisan agreement in congress with regards to funding for computing research, is, to some extent, because the CRA has been very effective in making the case of the value that it brings to the economy and the society as a whole in return for a small amount of investment.

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

Glad you're doing it.

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

Yeah, there are a couple of things I could pull out. One is a paper that I wrote with Abraham Silberschatz and Inderpal Mumick on what we call the chronicle data model and this was published in PODS[3] and nobody paid attention to it, but the whole point of it was that there is often too much data coming at too fast a volume for you to be able to store it before you process it. So we developed a data model for dealing with it in an online manner with data streaming. About 5-7 years later the database community discovered data streaming and the name was streams and not chronicles, but I am proud of having been there first and first by several years. The other piece of work that I'm really proud of is the TAX paper[4], which is the algebra that underlay the TIMBER XML database. This is a paper that was rejected at all the major venues and we eventually published in DBPL and I just think that, of all of my work, is the piece that I find the most elegant and it underlay the entire TIMBER system that came afterwards.

Can you say a little more about what the central result of the paper was?

Yeah, the problem has to do with how do you do set-oriented processing for something like XML where you are going to deal with different fragments and fragments may have different shapes. So you don't

[3] H. V. Jagadish, Inderpal Singh Mumick, Abraham Silberschatz: View Maintenance Issues for the Chronicle Data Model. PODS 1995: 113-124

[4] H. V. Jagadish, Laks V. S. Lakshmanan, Divesh Srivastava, Keith Thompson: TAX: A Tree Algebra for XML. DBPL 2001: 149-164

have a set of uniform structures that you can deal with. Our basic solution was that every operator would, as its first step, have something that renders things uniform and afterwards you would apply the operator and then we can develop a set-oriented algebra. So that was the idea.

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

I would blog.

Oh! Well, you blogged recently[5].

That was one of the first times, and that was more of a community thing. I would blog more[6].

Good. We'll watch for a blog appearing soon on your webpage. If you could change one thing about yourself as a computer science researcher, what would it be?

I wish I were better trained.

What an indictment of Stanford!

You know I got my degree in Electrical Engineering.

Yeah… but there were computer scientists in Electrical Engineering too.

Yeah, this is not an indictment of the university at all. In any case times change, things change and what we need to know changes. Its just that I seem to come up against the limits of what I know how to do all the time. I wish I knew how to do *X* and if I just knew how to do *X*, I would be in so much a better place to address some problem. Then I say "Well, I much teach myself *X*", and of course I never get the time to teach myself *X* and so, that's how it goes.

What are some example X's that you wished you knew more about?

I wish I were a better theoretician.

Oh, more theory! Okay. Anything else comes to mind?

I wish I were a better systems builder.

Woah! We covered both sides right there, okay. Well thanks very much for talking with me today.

Thanks Marianne!

[5] H V Jagadish. Big Data: it's not just the analytics. ACM SIGMOD Blog. http://wp.sigmod.org/?p=430

[6] Jagadish's blog is available at http://www.bigdatadialog.com/