Susan Davidson Speaks Out on Collaborating with Other Research Areas and Balancing Work and Family

Marianne Winslett and Vanessa Braganholo

Susan Davidson
https://www.cis.upenn.edu/~susan/

Welcome to ACM SIGMOD Records series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we're at the 2017 SIGMOD and PODS conference in Chicago. I have with me Susan Davidson, who's a professor at the University of Pennsylvania. Sue is an ACM Fellow, a Corresponding Fellow of the Royal Society of Edinburgh, and the recipient of the 2017 IEEE Technical Committee on Data Engineering Impact Award. Sue has served as the chair of her department, Deputy Dean of the School of Engineering and Applied Science, a member of the NSF Cise Advisory Committee, and as the chair of the Computing Research Association. Sue's Ph.D. is from Princeton University. So, Sue, welcome.

Thank you, Marianne.

There've been many decades of work on data provenance, but no one uses it! Scientists care about the provenance of their data, but they seem to be hand drilling their own solutions instead of using ours. Does that mean that we are solving the wrong problems?

I think that our solutions actually are beginning to be used, and can give a few examples. First of all, Zachary Ives' Orchestra System (Collaborative Data Sharing) is based on provenance. A fundamental aspect is the use of provenance tokens to evaluate trust. I've also heard that Boris Glavic has used provenance in his work with Oracle. And Laura Haas, in her keynote at ICDE 2017, showed how provenance was being captured and used in the context of the Accelerated Discovery Lab work.

So, I think there is impact. Where I don't think the impact is showing up is in a bench biologist's lab. And here I think the reason is a natural aversion to technology. Many people who go into biology don't really like technology or math. They really don't want to have to learn another workflow system that does automatic provenance capture. And they don't record provenance in the way that we think of it -- they typically use file names and notes to be able to record provenance.

That's cute that you started with biologists, but they may be the last to get on board.

They're late adopters sometimes.

What is the relationship between data provenance and data citation?

There's definitely a connection. Both of them are forms of annotation on data. Provenance as annotation is a big theme. Citation is annotation as well, but there's something additional in the snippets of information that you want to capture that may not be exactly provenance.

For example, you want to be able to record snippets of information that lets the reader of the citation know whether or not they want to look at the material being cited. Typically this involves something like authorship or title.

Can you give us an example?

Oh, yes. As an example, there's Einstein's famous paper on special relativity[1]. If you just give the reference to it, which is the journal in which it occurred, the volume, the number, the pages even, most people do not know that this is Einstein's paper on special relativity.

The same is true for the famous Watson and Crick paper on the helical structure of DNA[2]. Most people won't recognize the journal, the volume, the year. And so, additional information about the authorship and the title really gives people the intuition of whether they want to go look at that piece of work.

So, would you define data citation as provenance information plus a little marketing blurb?

Well, that's interesting, I hadn't thought of it that way. Certainly, it is provenance, but it may not be the "deep" provenance we think about, for example, in the provenance semirings work of Val Tannen. Data provenance typically starts from the very creation of the data as it was input to the database, and is tracked through queries, building up very complex provenance polynomials. The same happens in workflows, where data is tracked as the inputs and the outputs of processing steps that the data goes through.

In data citation, very often you just want to know where the data came from. Tracking back to see the influences of a previous work would be done through something like a citation graph, extending transitively past the immediate references. But we don't typically include that information in a citation.

And what makes data citation hard as a research problem?

I think what makes it hard is that the data is in a database rather than being published as an encyclopedia or a compendium of some sort. And the content in the database has been potentially contributed by many different individuals.

So, depending on the part of the database that you're interested in, the people who you should acknowledge are different. And there are a whole bunch of different parts of the database and a potentially infinite number of queries that would bring back a part of the database. So, you can't possibly attach an explicit citation to every possible query.

What you have to be able to do is to work with a small set of citations that the owners of the database attach to

[1] A. Einstein: On the Electrodynamics of Moving Bodies. Annalen der Physik 17: 891-921 (1905).

[2] J. D. Watson, F. H. C. Crick: Molecular Structure of Nucleic Acids: a structure for deoxyribose nucleic acid. Nature 171: 737-738 (1953).

pieces of the database, and use these to construct citations to general queries over the database. So, it's really the granularity of citations and number of different queries that makes this an interesting problem.

Being practical, can data provenance and data citation help solve the problem of fake news?

That's a really interesting question. Fundamentally, the problem with fake news is that you don't have a trustworthy origin for the news. So, if somebody claims that something is true without a reference to something that's trustworthy, then how do you know whether it's true or not?

If the person would give a reference, that is, they would provide the provenance for that remark, then you could go back and further evaluate whether you trust that as a source for that type of information.

So, I think it has something to do with provenance, and certainly, if there's provenance back to a trustworthy source, then it could be helpful. But in the absence of an endpoint which is trustworthy, there's really not much you can do about it.

[…] do you have a comfortable relationship with the person in that other field of whom you can ask stupid questions?

It seems like data researchers have gotten better over the years at picking problems to work on that are more likely to have an impact. Often that means looking at the data problems in a particular application area like bioinformatics in your case and recently air-conditioning, of all things in my own group. How do you do research that helps biologists or the air-conditioning industry when you aren't an expert on the subject matter?

Well, I've never worked in the air-conditioning industry, but I can talk about bioinformatics. I think that what it really boils down to is, do you have a comfortable relationship with the person in that other field of whom you can ask stupid questions? And do you have a basis – a vocabulary – with which to speak?

I like to think of this as a string of people holding hands, with different strengths and experiences. Somebody who can talk to the end-user who can also talk to somebody more on the systems building side. Maybe that person would talk to someone more on the theory side.

Not every member of the team needs to be an expert in all areas, in my case, in bioinformatics. And the biologist may not understand or even care about the technical solution. It's a team of people working together, talking – a lot of conversations have to happen. Frequently postdocs are really helpful, because they have the experience that a graduate student might not have, and they have the time that a professor doesn't have.

And how can you tell when you found the right team of collaborators?

I think it really is chemistry. A lot of this boils down to: Is this a group of people that you feel comfortable talking with, because you will inevitably display your ignorance. And if you're feeling nervous about somebody finding out what you don't know, then it's very difficult to ask the right questions and gain the experience that you need to come up with solutions.

How did you yourself become interested in bioinformatics?

I grew up in an academic family. My father was a professor of applied math, and my mother was a professor of plant science. So my mother was on the bio side, and my father was on the math side. When I went to school as an undergraduate at Cornell University, my sister, Jenny, was also there. She was three years older than I was, and she was studying biochemistry.

Jenny thought that we should take a course together since we were at the same university. And so I said, whatever you want to take is fine by me, and she said something really profound for the time (this was 1976). Jenny said that the future of biochemistry was computational and that we should, therefore, take a computing course. Computing courses were not popular back then, but she had the insight that it would be important. So, we took an introductory programming course – and I got hooked! I think our conversation gave me an appreciation for what computing could do to disciplines other than my own, and, of course, I have an affinity to biology because of my mother and my sister.

And you were a math major, right?

I started out as a piano performance major at Cornell, and realized by the end of the first semester that it wasn't a good major for me. I would walk into a piano lesson, sit down and start crying because I knew I was going to be crying by the end. So I focused on math,

which was my second major, and really enjoyed it -- right up until I took a very abstract course in topology, which I found very difficult to follow because I couldn't visualize it. One day, the professor (who had been scribbling madly on the board writing huge equations) walked over to the window, threw it open, and started barking at a dog. I didn't want to end up like that. So I thought I better choose another major. Computer Science seemed like a good one.

Was he barking at a dog or just barking like a dog?

There was a dog out there, and he was barking like a dog at a dog.

I would do that. I'm famous for doing that. I can work a dog into a frenzy with my bark.

Oh my gosh.

So, you switched to the wrong field, I think.

That's funny.

You've been involved with the Computing Research Association for a long time. What CRA accomplishments are you most proud of?

One of the reasons I've loved working on the Board of Directors of the CRA is that the people on the board are very service-oriented. They truly love computing research, and want to give back to their community. I like this type of person, and I like meeting them outside of my own field of databases. The other thing is that there's a disproportionately large number of women who serve on the board, and I enjoy being able to meet more women in computing.

One of the things that the Computing Research Association is well known for, of course, is the Taulbee Survey, which is widely used by departments for hiring and salary information. The Government Affairs Committee is also crucial, especially in these days when funding for science is becoming more difficult. The advocacy work that the Government Affairs Committee does is really important.

But the CRA also comes out with a number of statements and studies that can be used by the community. The most recent one that I was involved in was about the booming enrollments in computer science, a phenomena that has spread across the country -- the 2x-6x number of students in our courses, and the vast increase in the percentage of non-majors taking our courses. We did a survey, measured what the effect was, and tried to document how institutions were coping. As a result, we produced a report. That report[3] can be used by departments across the country that are trying to argue for more resources because of what they're facing in their enrollments. I think that this is really beneficial to the computer science community.

You had kids while you were still in graduate school. What advice can you offer for those who are trying to decide whether to start a family in grad school?

My advice for people is: start a family when you want to start a family, when it is the right time for you psychologically. The career issues will work themselves out. I chose to start in graduate school, which was risky because I was interviewing when I was pregnant. It was an awkward position to be in.

I started my first job as an assistant professor with a wee baby, which was extremely tiring. But I wanted to have children then, and I did, and I'm very glad that I did.

A lot of women wait until they get tenure, at which point fertility may be an issue; and you might never forgive yourself for not having had a child. So, for me, I preferred to take the risk with my career to regretting having waited.

My advice for people is: start a family when you want to start a family, when the time is right for you psychologically.

Our readers have requested tips from you on handling the balance between work and family life.

Always a difficult one. Interestingly, over the years I've had that question from as many men as women. I have always been very jealous of my nights and weekends, especially when my children were young. So I would work like a maniac during the day, and when I went home, I was with my family, with my children.

[3] Generation CS: CS Undergraduate Enrollments Surge Since 2006" by the CRA Enrollment Committee Institution Subgroup. Available at https://cra.org/data/generation-cs/

Sometimes I'd wake up in the middle of the night and start working, which my graduate students always enjoyed because 2:00 in the morning was when they were still up. So we were both up at 2:00 in the morning and could work on things together (remotely of course).

And weekends also. It is really important to be able to be at events for your children and do things with them. So I've always tried to be very efficient during the workday. I didn't spend a lot of time talking or lollygagging. I was quite focused on getting things done, and for me, that worked.

You've thought a lot about how to engage more women in computer science, including setting up such a program at Penn. What strategies have you found that seem to work well and others might want to use?

First of all, there are a lot of resources out there that we can use as departments, from the NCWIT, Women in Technology organization, to the Computer Research Association CRAW, a subcommittee of the CRA, to the annual Grace Hopper Celebration. There are also all sorts of resources that you can use to get students involved in undergraduate research, which I think is especially important for women.

The strategy that we've been using at Penn is to create a sense of community, so that women don't feel like they're the only ones dealing with the issues that they're struggling with. So we started a pre-orientation program for women coming into Penn so that they can come to campus ahead of time and get to know each other. We set up social meetings during the semester so that they can keep in touch. We provide options for them to be able to give back to the community by going to high schools, talking about computer science and how exciting it is.

We've also adopted strategies in how we teach computer science that seem to be more women-friendly. Peer programming and the ability to collaborate over homework assignments rather than working on them in isolation seems to very appealing to the women. And we've also tried to include in our courses as well as in our outreach events, an understanding of how computer science impacts everyday life. That computer science is not just a nerdish activity, but that it enables all sorts of good things, like discovery in medicine.

What is Dancing with the Professors?

At Penn, there's a Latin and Ballroom Dance student group that engages with faculty by having a competition each year, where they match up a faculty member with one of their student members. And you come up with a two to three-minute dance routine that you perform at the end of the semester.

It's just like Dancing with the Stars, but instead of a celebrity you've got a professor. I decided that I wanted to do it because I have always wanted to learn how to dance, and there's nothing like being given a deadline to force you to learn something.

So, I signed up for it. At the time, I was the Deputy Dean of the Engineering School, and when my Dean found out he was rather negative. He said, "Sue, it's very unprofessional, you know, dancing as a Deputy Dean." But I disagreed. I said, "I think it shows that I'm engaged with the students and that I want to be involved."

I saw this as a challenge, and really enjoyed it because it pushed me way past my comfort zone. I can get up and talk in front of hundreds of people, and it is not an issue for me. But memorizing a three-minute dance routine and performing it in front of 50 people was absolutely terrifying.

We've also adopted strategies in how we teach computer science that seem to be more women-friendly. Peer programming and the ability to collaborate over homework assignments rather than working on them in isolation seems to very appealing to the women.

What dance did you guys do?

We did a swing dance to "Shake a Tail Feather".

Aw, piece of cake, right?

You guys are good?

Do you know how I met my husband?

No.

Ballroom dancing.

Really?

Yeah.

Oh, that's great.

How does sports fit into your life?

When I was growing up, I didn't do any sports at all. But when I started graduate work at Princeton, I needed an escape valve for the pressure that I felt in pursuing my studies. So I took up running, and that has continued pretty much throughout my adult life.

It's always been some sport or other. It's either swimming, biking, running, yoga, strength building, or dancing. I've even taken up flying airplanes, which I don't think of a sport -- it was another crazy thing to try. But I think it's just to relieve some of the pressure that you feel when you're juggling so many different concerns between family and career.

Did you have a problem with injuries?

Only now that I've gotten older. Certainly not when I was younger. The warranty on my body parts expired when I turned 50!

I thought maybe that was why you switched from one to the other over time.

From one sport to the other?

Mm-hmm.

No, I think it's because I have a short attention span. Actually, my favorite sport was sprint triathlons because it's about a half-hour each (running, biking, swimming). You get to do something different every half hour, which is really good.

Do you have any words of advice for fledgling or mid-career database researchers?

The one bit of advice that my father gave me when I was young was: "don't think about it, just do it." And for me, that's been tremendously helpful. If I think about something for too long, very often I can convince myself that I shouldn't do it. Whereas if you go ahead optimistically and do your level best, very often you are successful. Fear of failure is common, especially with women, and prevents you from trying things. But sometimes, even failure is a good thing, and you can learn from it.

So, whether it's a paper that has been rejected from a conference, or a proposal that wasn't funded, or a student who decides they don't want to keep working with you, shake it off and keep going rather than getting depressed about it. I think that the benefit of age is that you've seen that in the past these things have worked out. The acceptance or rejection of a paper or proposal is a bit of a crapshoot. It's not necessarily an indication of the real worth of the idea or of as you as a person. You have to learn from failures as well as from successes.

Okay, so don't overthink it. Is that useful advice for daily life also? Job choice, shopping?

Shopping I can talk about.

Okay.

Usually, when you go shopping, you find something, and you instinctively like it or not. And then you think about it too much and end up walking away -- but then you return the next day to buy it. So I think that overthinking is something that we frequently fall prey to. I mean, look, we're computer scientists, we're analytical. We have to think about things, but overthinking is definitely a trap that we fall into.

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

I think that the work I did in workflow provenance with my postdoc Sarah Cohen-Boulakia is one of my favorites, because we started with real questions that people asking in the scientific community. We developed a beautiful formalism around it, which led to two Ph.D. theses, one by Zhuowei Bao and the other by Sudeepa Roy. Both had topics in their dissertation that were based on ideas from workflow provenance.

It was a very fertile field of work. One of my favorite papers (with Sarah) was on how to "zoom in and out" of provenance. How to abstract out from the details of provenance so that you can get an overview of it, and then how to dive in and see the details. This was a paper that was rejected from both SIGMOD and VLDB, and eventually published in ICDE. It's one of my favorite papers, and I think that it has had a lot of impact. This underscores my point of not taking failures too seriously. Have confidence in what you've done!

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

If I had more time, I would like to spend time talking with more people across campus about the problems that they're facing related to information gathering, data

management[4], and data analysis. I'd like to understand real problems in areas like sociology, economics, history, law, public policy and all the rest. I'd love to be able to talk to more people but really, it's a question of bandwidth.

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

I would like to be more intellectually curious about other areas in computer science. I would like to be more up on the advances in technology.

But as Department Chair, didn't you have to know all that stuff?

You do. You have to be aware of what the contributions are that your faculty members have made. But I would really like to take the time to go back and deeply understand areas like statistics, machine learning and data mining.

I've always felt that I didn't had the cycles to do this. I know that many of my colleagues manage to make the time, and I think I need to start doing that as well.

Well, thanks very much for talking with us today.

It's been great. Thank you, Marianne.

[4] Editor's note: this is now widely known as "Data Science", but was not when this interview took place.