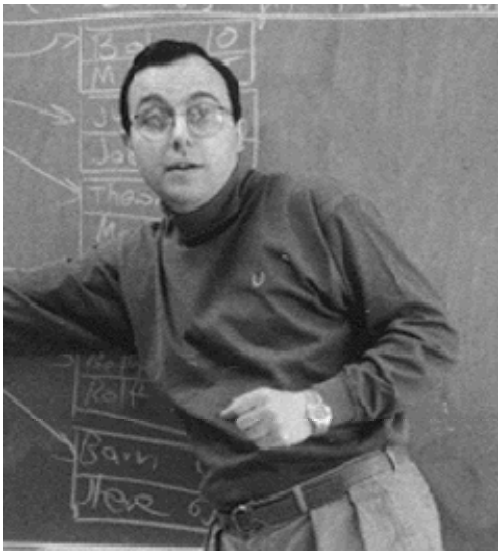
Yannis Ioannidis Speaks Out

on Database Research Funding in Europe, the Importance of Being Uncertain, Teaching as Show Business, the History of Histograms, and More

by Marianne Winslett



http://cgi.di.uoa.gr/~yannis/

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 2006 conference in Chicago. I have here with me Yannis Ioannidis, who is a professor of informatics and telecommunication at the University of Athens. Before that, he was a professor for many years at the University of Wisconsin in Madison. Yannis's research focuses on query optimization, digital libraries, and management of scientific data. Yannis is an ACM Fellow, has won numerous teaching awards, and received the VLDB 10 Year Best Paper Award. He is the Vice-Chair of SIGMOD, and his PhD is from the University of California at Berkeley. So, Yannis, welcome!

Thank you.

Yannis, your VLDB Ten Year Best Paper Award was for your 1993 paper, "Universality of Serial Histograms." Your paper was one of the first to introduce histograms in a principled way. How do people use histograms in the database world?

They use them to estimate how data values are distributed inside the database, and based on that, figure out which way to process queries. In the beginning, database systems had very rudimentary histograms, just assumptions that data values were uniformly distributed. With more sophisticated histogram, query optimizers have become much more accurate---not completely accurate in all cases, but much more accurate. And now all major commercial database systems use histograms in one form or another.

What were the research challenges in histograms for query optimization, and what influence has your paper had?

Finding what histogram to use was the biggest research problem. In the beginning, when people moved away from the uniformity assumption, they used the so-called "equi-width" histograms. And then they talked about "equi-depth" histograms, which were much better. And that was pretty much it. The main influence of my original 1993 paper was to show that there is a whole family of other histograms that will give much better results in query optimization, and are much more accurate, for many distributions of data values. And then all the other papers that followed that, both from my group or from many other researchers, really nailed the problem of what approximation is the best approximation for histograms.

Are there any new data management challenges that the database research community needs to be aware of?

The sky is the limit! There is so much that the community could be doing that it is not doing. I think we will be busy for many years to come.

One challenge is uncertainty. In the relational world where we grew up, data and queries are well structured and pretty well understood. You can do lots of things with relational data, and you know that you will get them right. Most of the world is not like that. We need to provide data management techniques that can handle uncertainty in the usage of data, in the accuracy of data, the provenance of data, and so on.

A second challenge is to serve all the other communities. There is a tremendous number of applications with huge data management problems, and we are not paying enough attention to them. I think we should. For example, we should consider how to deal with scientific data; how to deal with medical data and biological data; how to mix varied forms of data, text, video, images, and unstructured data. This is a huge challenge that we should be diving into.

Funding is a good carrot to get database researchers to go in new directions. Now that research funding is so tight for mainstream database research, I see more and more people teaming up with scientists. Certainly at every school in the US midwest, people are working with scientists and getting money through new avenues. What about the multimedia direction that you mentioned---handling different types of content?

Multimedia is at the heart of many of these applications. You cannot do many forms of science unless you have images from satellites. Much of medicine involves imaging and clinical data. All this multimedia comes to your table to be studied with all these challenges. As long as you try to serve other communities, you will hit the multimedia issues right away.

How hard was it to leave your flourishing career in the US and start over from scratch?

It was very hard. I was a member of what was considered the number one database group at the time, in Wisconsin. It was very hard to say yes to my family and to myself as well, and go back to Greece. And, you said it exactly right, it was pretty much starting from scratch. The system is very different, you have to establish new labs, and get new people. Also, the funding schemes are different, and so you have to start thinking in a different way; you have to change problems, because different things are funded on the two sides of the ocean. So, it was hard, but worth it in the end.

Can you comment on how the funding is different on the two sides of the ocean for database research?

Yes, and I will try to be politically correct in my answer. In the US, if you have a bright idea, as an individual, you can go get a small grant and study your idea and see how far you can go. There is not very much money, but it is there for you to do this. In Europe, there is tons of money, but there is no avenue to explore basic very long term ideas. You have to team up with lots of other companies, universities, and institutions from many countries, have a grandiose scheme, which must have some plan even for commercial exploitation, and that is how you get money. And the basic research, to a large extent, is done under the covers. So that is the main difference. Also, on a second level, because of that, the kinds of problems that get funding are different. There is no database research funded in Europe. Maybe this is because the main companies for database technology, for database servers, are in the US. So in Europe, officially, I don't do database research, I do other things. And, of course, through my funding that I get for other things, I do also my database research!

So what kind of other things did you migrate to?

Digital libraries, for example. And I don't complain about that, it is a fascinating field. I get inspired a lot by the problems that I meet there, both to do digital library research per se, and to change my database research. I am also involved now in some e-health projects where I am trying to apply data management techniques to health systems. There's no right or wrong funding system, it is just different in different places. In moving to Greece, and to Europe in general, I had to change my focus.

Do you find that you spend your time differently now that you are a professor in Greece, versus how you spent your time in the US?

Yes, because of the different ways things work in Europe. I spend more time connecting with people, both to come up with ideas to write proposals, and also to do the work. The projects in Europe are a lot more administration-heavy, so I have to deal with that overhead. Teaching load is heavier, in Greece at least. I teach two courses every semester. Classes are much larger. The courses that I used to teach would have 60 undergraduate students in Madison, and in Greece I have 200 students. In the graduate courses in Madison I used to have 20 or 30 students, and in Greece I have 80 students. So, all of this has an effect to have less time to spend on research, or less personal time. Another thing I am trying to spend some time on in Greece is trying to change some things that I saw in the US and I liked. Together with some of my colleagues, we are trying to change people's attitude towards certain issues. So I am putting some time into, let's say, foundational changes---sometimes successfully, sometimes not successfully, but it's something I am doing.

There seems to be a bit of a reverse migration of database researchers back to their home countries in Europe. Have conditions changed in some way that makes this a more attractive route than it was before?

Absolutely. I believe the European Union has played a major role there. In the past 20 years, the EU started funding "research framework" programs. Technological areas, computer science in particular, have a constant stream of money going through these programs. So the ability to do research, or at least the funding for research and development in an academic environment, has increased and keeps increasing, and there is no sign of it diminishing. Since the opportunities are

there, it is much easier for someone like me to return, and with some effort, to rebuild and do partly, at least, what he or she was doing back in the US.

What do you think of "publish-or-perish"?

I believe it is wrong, when it comes to the quantity of publishing. Obviously you have to publish, nobody would remember you for ideas that you never spoke of or never wrote, so you have to publish. But, I would rather see someone with three good papers than someone with one good paper and 10 mediocre papers. Often people are judged first by the numbers, and then by the quality. Not always, but often. I would say it is better to spend more time on a problem and do it right, and have one really good paper, rather than the other way around.

I hear that you are a fantastic father. How do you reconcile the time commitments for a good family life and a good research career?

I am sure you haven't heard that from my children! I don't consider myself a fantastic father. I think I could be much better than I am. It is very hard for anyone to reconcile those time commitments. I'm trying to spend as much time with my children as possible. At least, if I don't spend as much time with them as I would like, I want the times that I am with them to be good times, educational times, and fun times. It is often hard now that I am in Europe and all these research projects have so many partners at remote sites. Plus I travel quite a bit more than I used to before, say 10 years ago. I am often away, and that is really tough on me and my kids as well. I used to buy tickets so that I would have a Saturday night stayover, I don't do that any more!

I have heard that you are a very hands-on advisor. With that advising style, how can the students be ready to "fly solo" when they leave your nest? And more generally, what have you found to work well in advising students?

I think it is accurate that I am "hands-on", but not in the sense that I hold them by the hands and guide them. I am just there next to them. In the beginning, I hold their hands, and then, little by little I let them go. But I am just there, I want to understand the details in order to teach them about the details when I see something wrong. So, in the end, I believe they will end up being stronger and readier to fly solo than if I had just thrown them in deep water and asked them to swim. I consider this a benefit as opposed to a disadvantage.

Being a friend in addition to an advisor has worked well for me in advising. Getting a PhD is an issue of self-discipline and growing as a person, as much as it is growing as a scientist. It is a lonely way with some ups and many downs. Until you get your first SIGMOD or VLDB paper, you have, more often than not, rejections in other places. Students can lose courage when their first attempt at a paper is not successful. So you have to be there to support them, to teach them patience. I think this is the most critical thing, next to the scientific things that you teach them; to be there with them and support them psychologically.

I have heard that you enjoy being on stage, and were involved in the theater in Greece as a youngster. How does that affect your teaching and speaking style?

Very much so. I think the way I teach is a 100% influence of my theatrical experience when I was in high school. When you are up on stage, you have to communicate with the audience, forget who you are, immerse yourself into your role, and then give this to the audience. I think that is the way I teach. I forget who I am and I immerse myself into the topic that I am supposed to teach, and I try to communicate that. However, in the theater, there is not much interaction with

the audience, other than that you are looking at them. In some modern theater you may have more interaction, but in classical theater you don't. When teaching, you have the additional advantage of being able to explicitly interact with the audience, and then sense how things are going, and proceed accordingly to convey that. I think teaching is show business, just like theater. I became a different person after my theatrical experience than before.

You were an advisor to the Minister of Health in Greece. What was that like?

A very different experience. I was called to be an advisor because of my expertise on data management and computer science in general. On 5% of the cases, I used that expertise, and then the remaining 95% was completely new things: how to manage large organizations like the ministry, how to say what you need to say without offending people because of politics. There was a lot of public relations work, too---not that I did all this, but I had to face all this, so it was a great learning experience. I don't know if I would ever do that again, but I learned a lot.

Often, as researchers, we have the greatest ideas and we come up with the right solutions, and then we try to apply them to reality. But then the human factor comes in, and issues that you never thought of are raised as problems, and resistance arises to applying these wonderful new things that could change the world. You don't know what to do, you speak a very different language from the rest of the people. Often I felt that living in academia or in computer science in general, we are in our own little world. We think that the rest of the world operates like us, but it does not. It is a wild world out there! It was fun, but I hope it was the first and the last time I did that.

Your interests in scientific data management have led you to work with real scientists. How was that experience?

Very interesting. It is trying to interact with a different world. It takes a long time for the two communities, computer scientists and people from other sciences, to understand each other, to start speaking the same language. So it is a lot of effort up front, but it is very rewarding. In high school and somewhat in college, you learn about rudimentary things from other sciences, and then you forget about them because you dive into your own science. Working with scientists gives the opportunity to bring back all the subjects that you used to like back in high school - for me biology, astronomy, physics, and other things. It gives the opportunity to learn some more about them, and also, to a certain extent, see what we do in computer science be useful. Often when you work in the internals of servers you never see the results of that work affecting people. A few cases where I saw that were extremely satisfying. I recommend it to all of us.

At Wisconsin, the computer science department gives out a teaching award, and the students also choose a recipient of a teaching award. You won both awards so many times that I heard that they made a special award just for you. I have heard that you are especially good at teaching database theory to systems students. How do you do that?

Indeed there was a special award just for me when I was leaving Wisconsin. They gave me a lifetime award, which was really nice of them. It is very hard to say what it is that I do that makes my teaching style successful. Here are a couple of things that I can think of. The first is the interactive style that I mentioned before. Often I stop, I ask questions, I ask people to vote on ideas, I ask them to give me answers, then the class votes on those answers, and this keeps a good percentage of the audience engaged. I think this plays a major role. The other thing is that often, I don't just teach what it is, but I try to bring it out from them. I lead them along the way, I ask

them questions, and wait for responses from them. And if they say it, or one of them says it in a way that they all understand, it stays in their minds.

So how do you do that now that your class size is 300 instead of 30?

The same old way. I don't use PowerPoint. In undergraduate courses, I still teach on the board. The pace is slower so people can follow more easily. I often try to lead the class into a wrong answer to a question, and then bring them back and tell them what the right thing is, and I ask them not to remove it from their notes. I want them to have gone through the wrong, and then gone back to the right answer.

Don't the students complain vigorously if you don't give them PowerPoint?

No, not really. I haven't heard that complaint. Actually, I teach a human-computer interaction course in Greece, I didn't use to teach that in Madison. The past couple of years, it has been the only course that I started teaching with PowerPoint, because the topic is user interfaces, so it is much easier if the students can see the interfaces. And a few people complained. Okay, they said, use PowerPoint for the interfaces themselves, but do rest on the board.

At Illinois, if you put anything on the board, the students complain that they can't see it, because the class is so large that they can't really see the board. PowerPoint is projected very large, so everybody can see what you are doing.

In the large classes in Athens, we use a classroom where we videotape the lecture. We videotape the board, and these videos are projected on screens for the students to see. And also, you can watch from your house over webcast.

David DeWitt says that we need a new paradigm for query optimization. Do you agree? What should that paradigm look like?

I don't know if we need a new paradigm, but certainly there are lots of things we haven't solved, and we need to study them. If the same paradigm works, that is fine; otherwise we need a new one. The key thing is uncertainty: mixing optimization and query processing; optimizing for different kinds of answers, not accurate answers; optimizing not for speed but for completeness or for freshness of data; optimizing in a distributed system when you have autonomous systems that can give you some answers but you cannot control them. If I wanted to put these kinds of different environments under an umbrella, I would call them uncertain environments. They are certainly a big challenge and definitely may well need a different paradigm than normal.

You are interested in human-computer interaction (HCI) issues, which the database community has traditionally been very poor at addressing. Is there new hope for us?

I think there is hope for us if we put our minds to it. The HCI community has made tremendous progress in general with user interfaces. Database querying and database interaction are a special form of user interfaces, and may have particular difficulties that have caused us to make little progress. We can never be sure, because really we haven't spent the time necessary to know. For easy queries, we did QBE, we did query by forms, and now the web works like that. But how do you express an aggregate, how do you express nested queries? It is tough, but we haven't spent enough time working on it in order to crack it.

Should database systems people care about schema equivalence?

Stonebraker says no, only 5% of the problem is schema equivalence. I would say yes, but in a broader sense, not purely schema to schema. Most of the world is not schema based, but still you need to have mappings. Let us put all the items of the entities (those that are in the data, those that are in the schema, and anything else) into one big bag, and then see how to find equivalence across the bags.

You did research on schema equivalence in the context of the extended relational data model. What new challenges arise when we want to reason about XML schemas?

I don't think any new brand new challenges arise. It is just that the problems are harder because in the relational world you have more structure. In the XML world, you lose some of that, and then how do you map across? Some people have done some really nice work on that.

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

Think out of the box. Address challenges that nobody else is looking at. And, especially, what I said before, find challenges that other scientists, or other parts of society, have. Try to address those problems. There are many problems to be solved there. The impact would be larger in the end. I mentioned a few such areas earlier, let me not repeat them here---but go out and collaborate with other sciences. Some of us must still work on server stuff and classical problems that reoccur in new environments. But the great majority, especially fledgling and midcareer people, should go further away and dig there. There is a lot of gold there.

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

Let's see, which child of mine do I love the most?! Well, certainly, the histogram work has a special place in my heart. It was exciting and very different. Also, it has had the biggest impact of all the work I've done. But even before this was realized, it was a problem that people could have looked at many years before I did, but no one had. It was well defined, and it had a very interesting solution. So I love that work. Also, I like my work on scientific databases. It hasn't had as much impact, but I believe it is very important in the systems that my team and my colleagues build now based on the results that have come out of that. I am also excited about personalization, which is my latest project.

What challenges do you see in developing user preference models?

User preference models and personalization in general is a field where you have to move away from the machine and bring the human into the picture. And then everything breaks loose. Humans are unpredictable. We don't know exactly how humans operate. So coming up with models of preferences, how people think about preferences, is extremely challenging. Yes, I believe there is hope. There are some models, or some partial models, that seem to work well. Some of them are useful in databases and data management in general. Some others are for other types of applications. It requires, at some level at least, interaction between computer science and psychology, which makes it very exciting.

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

I would probably work on user interfaces for databases. I think there are things that can be done there. There are many other things I would like to do at work, but I think that one would be at the top.

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

For some things, I wish I knew more mathematics. It would help push certain areas deeper than I have been able to push them. And speaking of pushing, one thing that I would change in myself is the following. When you work on a problem, you work on it for a little while, and then you move on to the next problem. In certain cases, I have moved to the next problem too soon. I believe I should have stayed and pushed a bit more.

In hindsight, where would you have stayed longer?

I would stay longer on heterogeneity and schema equivalence, things like this. I would have stayed on user interfaces, I've done a little bit of work, but I haven't stayed long enough.

Thank you very much for talking with us today.

Thank you.