Jiawei Han Speaks Out

On data mining, privacy issues and managing students

by Marianne Winslett and Vanessa Braganholo

Jiawei Han
http://www.cs.uiuc.edu/~hanj/

Welcome to ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we are at the University of Illinois at Urbana-Champaign. I have here with me Jiawei Han, who is a professor of the Computer Science Department at the University of Illinois. Before joining Illinois, Jiawei was a professor at Simon Fraser University in Canada for many years, and briefly before that, he was a professor at Northwestern University. Jiawei's research interests lie in data mining. He is editor in chief of ACM Transactions on Knowledge Discovery from Data and he is the coauthor, with Michelline Kamber, of a popular textbook on data mining. Jiawei is an ACM Fellow and an IEEE Fellow. His PhD is from the University of Wisconsin at Madison. So, Jiawei, welcome!

Thank you! Thank you Marianne!

So Jiawei, according to Publish or Perish, you have an H-index of 76, which means that you have written 76 papers that have been cited 76 or more times. To put that in perspective, Publish

or Perish lists my own H-index as 27, and I'm pretty happy with that. But of course I'd be even happier if I were three times as productive and also had an H-index of 76. So tell us, what advice do you have for those of us who would like to be more influential?

Actually, I myself think this is the first time I have heard about 76[1]! I remember when somebody told me there is an H-index… I went down there, I looked at it, it was around 54-55 something. I never knew I got 76. I think probably the best thing I can say is: if you choose a topic, choose something pretty exciting. Maybe it is tough, or maybe it may not be that tough, but I think, choose something which is a little fresh, a little meaningful, a little exciting, and try to find some good solutions. Probably, that is the best, because usually you get a paper you like people to read, and that you yourself like to work on. So if you got excited, it's likely that other people might get excited as well. Of course, not every paper can do things like that. Of course for me, if I got more papers published, I'd probably prefer every paper I write, just not write that many, but just write something more exciting. But since I am also supervising lots of students, you cannot expect every student here to write every paper not only exciting to the students themselves, but exciting to everybody. That's very hard. So we do have some papers, maybe a little incremental, that may not be so exciting. But I think overall, work with more exciting topics, and also try to work out some neat and elegant solutions, then probably people will feel more exciting to cite your paper.

So let's talk about data mining. Data mining is very important and popular today, but it is a very young field, with the first paper appearing only around 1992. What led you and the other KDD founders to start working in that new area?

So a KDD paper… If you say it appeared around 1992, it was probably in a database conference, like the SIGMOD or VLDB conferences. But even in other conferences, I probably can trace back to 1989, when Piatetsky-Shapiro organized the First International Workshop on Knowledge Discovery in Databases. That one I still remember because I sent a paper there, I attended the workshop. At that time it was a very small group, about 30 people. It was actually attached to IJCAI, that was in Detroit. At that time everybody felt this could be a big fish, a big direction. I myself also felt this way because I worked, as my PhD thesis, I actually worked on deductive databases. At that time, logic programming, and database was very hot.

I got a big influence by Randy Katz. Randy at that time was a professor in Wisconsin. He actually once gave a seminar. At that time, Japan got a 5th generation computer project. That was 1983 or 1982, I forgot. He actually, in the seminar, he even put a Japanese sword right on the table. He said it was a Japanese challenge. So he said the Japanese wanted to work out a Prolog machine, which is highly parallel, that can do a lot of database searching, inferencing. He said we needed to face the challenge. So I got a deep impression on that. That time also was the

[1] Editors' note: at the time this interview was published, his H-index was 101 (we used Publish or Perish with the search string "Jiawei Han" and area "Engineering, Computer Science, Mathematics").

time I selected my research topics. So, I got pretty excited, actually, once I got there. I did my thesis on that, published quite a few papers, including both database conferences or even logic programming conferences. So it is a rather different field, you will see resolution, you know, (Herbrand Universe), recursion, recursive query processing, compilation, all of those things probably in the other domain. And when I went to Simon Fraser, I was also interested in looking for good topics to extend the scope, because, I myself feel if we just rely on expert rules to derive knowledge, or new knowledge, it is far less efficient; you really need tools to dig through the data to get knowledge rather than just relying on expert rules. So that is the reason I am actually quite interested in integrating, like machine learning induction into the database as well. That's the part where we started.

We got a paper, I actually originally did not really know where to send, because we got the paper, and there were no such conferences, actually. Of course, we could send it to database conferences, but I did not really know whether people would like it. At that time, Gregory Piatetsky-Shapiro organized one. He just sent emails that said he was organizing the first international workshop on knowledge discovery in databases. I figured this one actually was a very good match of the algorithm we worked out. At that time we called this as attribute-oriented induction. It works in our databases going up and down, then you can derive some generalized knowledge. So we sent the paper down there, it was taken very quickly, because it was a workshop. That was the first paper we got down there, so then we did lots of improvements, and extension, and later we got it into VLDB 1992[2], that is the one you mentioned.

In 1992 the database conferences started taking data mining papers. But the interesting thing is, I think there were quite a few milestone topics on data mining before the formal, like the KDD conference formed up. Actually, almost all appeared in database conferences. It was very highly cited, very excited thing on this. I think the reason could be this: I believe the database people like to work out the algorithm, really working on very huge amounts of data, scalable algorithms, and also they really worry if it's effective, you get all the performance time. That is the reason you'll probably see, I can give you a few good examples… Of course, Rakesh Agrawal, their associational mining paper was probably the most cited paper even in the database conference history, and Raghu Ramakrishnan, their paper on BIRCH, that one is also on clustering algorithms... Even Johannes Gehrke, on the Rainforest algorithm. There are lots of very highly cited papers that actually appeared first in database conferences. So to that extent, I should say, data mining actually grew, of course, you could say it actually grew out from many different places, from machine learning, statistics, database, but database conferences really took a lot of very good papers, those are milestone papers in the KDD history.

What application areas for data mining are just around the corner?

[2] Jiawei Han, Yandong Cai, Nick Cercone: Knowledge Discovery in Databases: An Attribute-Oriented Approach. VLDB 1992: 547-559

Oh, data mining, I should say, can almost apply to anywhere. You probably can see it. For example on web search, people use, say a PageRank or HITS algorithm. Essentially, PageRank or HITS is also doing data mining, because if you found a page that was pointed out by other ones, it really carries semantics, carries importance. That is the reason you finally can find very interesting, very relevant pages. So, my feeling is, the first thing we should think data mining is invisible data mining. That probably is the most interesting but most effective mining method. There are people using it. You think about Amazon.com, they say "People buying this book also buy other books". They are using for example collaborative filtering algorithm or some other data mining algorithms. You think about Google, people search on the web, they are using some mining results. So those invisible data mining, even if they do not say they are really doing data mining, actually they are using this methodology. I think this is probably the most interesting thing to see. There will be many things coming out.

Would you think someday that we will have domain-independent approaches to data mining, like a unified discriminative ranking model that's independent of semantic issues, independent of the particular application? For example, if you see a customer behaving in an unusual manner, maybe that means fraud, maybe it means that might be a big spender, or maybe it's just noise. Do you see domain-independent approaches to basic mining tasks?

Actually, for data mining as a discipline, you want to work out some general principles. To that extent, you don't want to stick with very, very concrete, you know, say, my methods are just working on this particular problem. You want to be a little general, somewhat domain independent. But on the other hand, because of different kinds of data, you really need different methods. They are so different. For example, just mining sequences. The sequence on transaction databases, like shopping sequences, and sequences on biological data, like DNA sequences, biological sequences. or mining text sequences. could be very, very different, because they are looking for very different patterns. So if you say, my algorithm works on all kinds of sequences, probably it is good for nothing, you know, it really cannot find patterns! To that extent, I think we can say the algorithm is first tailored to this particular application. Then you work out a very effective algorithm, maybe you try to extend your scope, to work on other applications which could be somewhat domain independent.

I remember, actually, we worked out one method I think called CloSpan, that's the one. We first worked out PrefixSpan by Jian Pei with me at Simon Fraser, and Xifeng worked out this CloSpan algorithm. I remember one professor in Purdue, he or she took this algorithm, and actually tried to use it on biodata, and also found something interesting. And I remember in our CS 591 seminar just a few weeks ago, there were some Japanese researchers, they do, I think it's more like web log, or web blog mining. They first used our PrefixSpan. I did not actually even realized, they first used this, so to that extent, this one can be used in multiple domains. So I like things to be more domain independent, but I think for particular questions, for particular problems, we have to first focus to make it more specialized, to make it work, then think how to generalize it.

Does that mean, does data mining have any general principles, like databases do?

I think the problem in database and data mining could be a little different, especially if you think about relational databases. The data is more like really well structured, and for this structured data, you can work on like selection, join, query processing or transaction management, reasonably easily. Actually if we just work on highly structured data, there could be some algorithm you can transfer to different domains . But even that, people are looking for different patterns. For the same structured data, you may look at different kinds, whether you want to find clusters, you want to find you know like regression, or evolution. Since people are finding different things, likely those algorithms will be tailored, or for different applications, they could be rather different. That's why, some people dream we have on-the-shelf, data mining tools. You just download it, and every pattern will be there. At least, at this point, I'd say, it is not realistic. It may not be quite effective. So you have to know the domain better, and you really know what's the pattern you want to find. and what's the trick you can use your knowledge. I think that's, it is far from like just using a very simple language, like SQL or SQL mining, you can solve all the problems.

Ok! So, two people suggested that I ask you about the ethics of data mining. One person gave me ChoicePoint as one example. I looked up ChoicePoint on Wikipedia, and it says, "ChoicePoint […] is a data aggregation company […] that acts as a private intelligence service to government and industry. […] ChoicePoint combines personal data sourced from multiple public and private databases for sale to the government and the private sector. The firm maintains more than 17 billion records of individuals and businesses, which it sells to an estimated 100,000 clients […] However, this data has not been secured sufficiently to prevent theft of data on at least one occasion. […] The company has also been the subject of lawsuits for maintaining inaccurate data, inquiries whether it allowed political bias to influence its performance of government contracts and accused of illegally selling the data of overseas citizens to the US Government." So of course there will always going to be mistakes and inaccuracies in mined information, and there will always be some people who are greedy or corruptible. If we look forward to the future, how can we address these problems for data mining?

Actually, I read a lot of newspapers or some different controversial things on data mining, so the first thing I should say is, for any research, for example, when you apply for an NSF grant, they will ask you "are dealing with human subjects or non-human subjects?". Data mining is actually dealing with both things. A lot of data mining things are not dealing with human subjects at all. For example, if you try to mine some astronomy pictures (like what Jim Gray did, it was astronomy databases), you still need a lot of data mining. You probably will never worry about disclosing any stars privacy. So to that extent, there's no real privacy issue. Actually data could be public to anywhere in the world, anybody can share. So there are lots of such data mining tasks which we do not have to worry about the privacy.

But on the other hand, there are human subjects. For example, you mine data related to the people. So definitely once we get into this one, we have to think about privacy and security, all those issues. What I feel is, with data mining, usually you can think in two ways to do data mining. One way is, you got an in-house data mining software, you mine by yourself. For example, Wal-Mart can have some data miners, they sit in the Wal-Mart database, they do all the data mining. Then, they may have a choice, what they should publish, what they do not publish. Even for the in-house data mining, there could be issues whether you could look at any personal data or not. So as long as you have some appropriate measures for in-house data mining, for example, you say, I mined a customer record. Even you can take, say a particular credit card number linked with a previous same credit card number, and you find the shopping sequences. As long as you say, this employee or this data miner has no way or is not permitted to look for further links from this number, then you treat this number as a dummy number, like RFID, so you finally find something. That way, based on my viewpoint, you still haven't violated anybody's privacy yet. But the problem mainly is, what things you can publish? For example, if you publish things in a more statistic term, for example, you see US Statistic Bureau. They regularly publish lots of things. You can buy a CD-ROM, and you can have many years of data. You can go down to zip code, but zip code is still quite big. In most cases, one zip code may cover say thousand, or tens of thousands of people. If you say K-anonymity, this K could be ten thousand. And there is no way you could build links, so you publish this kind of data is still safe.

Based on my view, if you do in-house data mining, and then you carefully publish your data, which make your K or make all these privacy preserving things quite big, you are reasonably safe. You still can use this data. I think many people are using it, like US Statistic Bureau, they publish lots of data, lots of people are using it. I believe these things are necessary because an administrator like Obama wants to know some concrete statistics. Anybody who makes decisions need to base it on your data. So for those kinds of privacy, if you make this K quite big, you should not worry too much.

But you can charge more for your products if you make K smaller.

Yes, that is exactly that. You get a little dangerous then. If you make it too small, people start identifying something sensitive, something that may really violate people's privacy. That's exactly why privacy preserving publishing and data mining actually becomes a very important topic because people want to do both, want to find deeper information, but in the meantime, protect people's privacy, and these two could be conflicting goals. But another very important thing is (some people discussing about this) out-of-source data mining. I feel this out-of-source data mining could be a little dangerous.

What is out of source data mining?

That means that you ship your data to other people to mine it. Then, the other people mine it, and you don't want this miner to dig up more information than you want. This thing, my feeling is, it

is very easy to go out of control. There are lots of research papers. My feeling is, some research papers say you can do k-anonymity, do l-diversity, do t-closeness, doing all these, and there is one professor from UT Austin who actually showed: if you very secure to guarantee this, then you do data mining, probably then you cannot really find good patterns. Actually, it could be even worse than the intruders or something. So I think this could be true. To completely make your data like have no characteristics, then ship it to other people to mine, you probably won't really find very interesting patterns. But on the hand, if you ship the raw data or some data that really contains sensitive information to a third party to mine, I don't feel it is a good idea.

One person commented that startups are much easier to do today, and wanted to know when you plan to start a company, and what your product will be.

The first thing, about the startup and whether it is easy to do or it's hard to do: different people may have different opinions or experience. My feeling is this: there are, of course, people working on database or data mining, a very practical domain. There could be lots of applications. Those applications may promote lots of startups. Whether researchers need to do startup or some other people may take the idea to do a startup, that completely, different people may have different opinions. For me, I actually like to concentrate on research. That's the reason I'm not that interested to set up a startup or something. For research-wise, I already think I got quite exhausted. If I do startup, I probably would have no time to sleep!

Have you found any challenges in your graduate students being distracted by industry or by startups?

I think for the students, actually being in UIUC at Urbana-Champaign is much better than in a big company or in a big city. For example, somewhere really in the center of the bay area or in Seattle or some places. Those places may attract the students in a very easy way because they just give a phone call, they just ride a bike, or drive a car, you can go there. I believe here, I do not feel the students here really are distracted by those companies. To a certain extent, it is good to learn something about a company's needs, to see the real world. I encourage students to go out to do summer intern, especially go to a real company, go to research labs, to do summer interns. I feel this is a very good practice, because you learn something about outside world, about applications, about industry, about research labs. When you come back, you probably have different research problems, different ideas, you build up your social network, research network. I think these all will help students.

With industry and academia working so closely now, what guidelines do you recommend for young graduates to choose between working in academia and industry?

I think students have different thinking, different preferences. Some students really like to go to industry. I got one student, Zheng Shao, who is very smart. Unfortunately, he did not finish the PhD. In the middle, he actually was attracted by Yahoo! first. He left, he went down there, he did very well, he is very successful. He actually came here I believe a few weeks ago doing some

recruiting. I think the students really liked to go to industry. I still like them to finish PhD, because once you get better knowledge, and you master the research area, and also you get into PhD, likely you get into a little more like a research or development, or more invention or some kind of position where you use your talent better. To that extent, I encourage every student to finish PhD before going outside.

But on the other hand, I know there are lots of students that really want to do research. I told them for doing research, basically you have two major choices, one is going to university, because doing research in a university is not just doing teaching, you actually really dash forward, by working with graduate students you can do a lot of very interesting research. And another one is actually is good industry research labs, like IBM Research, Microsoft Research, Yahoo! Research, Google Research. There are lots of such research institutes or research labs. I think those research centers are very exciting as well because they got a lot of people, they all have PhDs, with different talents, and they are usually really good. They work together. And they also can work with real industry, and work with a lot of professors who may go down there for sabbatical or for doing some joint work. You really can, to some extent, extend your scope. For example, like Xifeng Yan. He did probably two or three years at IBM Research as a researcher, then recently he joined UC Santa Barbara, actually as a Chair Assistant Professor. He got very good training on both academia and research labs.

You were a young person in China during the Cultural Revolution, and then suddenly you were a graduate student in computer science at Wisconsin! How did you make that transformation?

I think this could be a pretty heavy topic! So, it is true, at that time it was a very unusual transition, I should say. Not only I was young when the Cultural Revolution broke out, it was a pretty dark time, in the sense, my family was intellectual, so I was almost at the bottom of society to some extent. I labored in the countryside for quite a few years. It was not easy. Actually, not only me, the whole country, to a certain extent,,the university was closed for almost 12 years: 1966 to 1978. For me, myself, of course, it was not easy.

Probably the really breaking point was in early 1978. China restored the Graduate study, Graduate School. I was bold enough to try the Chinese Academy of Sciences, and I passed the exam, and also the English one. So, I probably could say 1979 was the first year China and the US got diplomatic relationship, and I went to the University of Wisconsin in 1979. So I was very fortunate, but also it was not easy. It was interesting because those many years, China, I should say, almost closed doors for 30 years. That was the first time actually people even went to campuses to see the students from China. Even Wisconsin I think was bold enough to take those students. The reason was at China at that time there was no TOEFL, no GRE. There was no exam system at all. But I remember that a University of Wisconsin Professor told me they were bold enough because they saw China was a very big country, and it is a very big country as well, and there must be many talented people, Actually, one professor who wrote recommendation letters for me, got a PhD in UIUC, I should say. But also very unfortunately, he was labeled as a

rightist and was almost forbidden to do any research work for 25 years. But he did write a letter for me and for Hongjun Lu. At that time, when we first came, Wisconsin actually first put these several students as international special. not formal graduate students, because they could not judge us. But they said, since the professors said these are the very best selected students, got into the Chinese Academy of Sciences graduate school, so they must be good. So they just blindly believed that. They took us, and we easily passed the first year, so then finally we got through that.

Actually, in those several years, there were many such cases. I remember YY's (Yuanyuan Zhou) advisor, Kai Li. YY actually invited him to come to give a talk, and she also invited both me and a few other professors to go down to her house. When Kai met me, he was joking, he said, "You know, we were classmates together with Ming Li," he says, "You see, that year, that two years, probably now only a few Chinese came out. Among probably four or five Chinese ACM fellows, we 3 are from the same class." That simply says, even we three people, it was just a tough competition, I probably should say, when you finally got there. I just say, China opened up, brought really a lot of changes for China itself, but also for lots of students.

You spent your sabbatical year (2007-8) in your office doing research and writing proposals. Why not get away and do something different instead?

I think with so many students working with me, it is a little hard—you are out of students' scope, going out for a very long time. And also there are so many things waiting for me to get done. I think for that sabbatical I did a lot of things on not only research but also enhanced my book, there were lots of other things. So, I just have very little time, hopefully in the future, I may get a chance.

I had a question suggestion I found very amusing. Are you ready?

Yes*!*

Which of the following would you find the most rewarding: (a) writing a landmark book in the field; (b) graduating lots of outstanding PhDs; (c) coming up with your own good algorithms; or (d) winning lots of prestigious awards?

That's a little hard to say, but I probably should say, training a lot of outstanding graduate students, that is probably the most exciting thing. The reason is, not only you train a lot of students, but also they will become the new seed for the whole field. If you don't train a good number of students, the whole field problems really cannot be solved by a small number of people. So I think that is really the most exciting thing.

How many graduate students, postdocs, and visitors do you have right now?

Okay, so, I think I can't roughly count, maybe. It depends on how we count it. I think I have 17 PhD students, and I have 2 master students, 1 visiting scholar so far, and 1 visiting student.

So, that's a great leading for my next question! Someone told me, "I am amazed by Jiawei's time management skills and his enthusiasm in research. Even with his large group, he still answers most students' emails within hours, no matter whether it is in the morning, noon or midnight." So, how do you do that, with so many?

I think, the first thing, for students supervision, with many students, there are disadvantages and advantages. The challenge could be how could you handle every student because everyone is different. But from my viewpoint, you should not do everything just by yourself. The students can work together. Another thing is, the students can organize by themselves as well. So in many cases, no matter in Canada or in US, I feel there are always, there are some students, who really have the talent, who can lead, who can be the future, who will probably be a leader or professor, so they can really organize things. You should really let those students to play some role. In the meantime, I meet many students actually in groups. To some extent, these groups are dynamic, in the sense that, once we have some research topic, sometimes I send an email and say "who is interested in these topics?". There will always be a few people volunteering, and some are very enthusiastic. I will ask those enthusiastic students to lead, and then we form research groups. We finish this, or even before finishing this, there could be some new topics, some student, very energetic can join 3 or 4 different research groups. I think this probably can really help reduce my load.

But for answering email so quick, I think the first thing is I try to, sometimes try to forget, otherwise email will pile up, and I have to read it again and again. It may not be a very good habit. I think Johannes Gehrke actually, I remember once in SIGMOD Record he wrote an article, he said he tried not to be distracted by emails. He tried to pile up the email until a certain time, like 3 o'clock. I think that could be a good habit, because you really can concentrate more on your research. Sometimes I could be distracted.

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

Yeah! I think one thing probably is choosing promising and good research topics. Usually, I should say the midcareer researcher should be bold enough to challenge some new things. That's the one, see, I started working on data mining to some extent to challenge myself. If I feel this is a very good topic, I would like to jump in. I think for midcareer researcher, especially if you already got tenure, you really should be bold enough to find something you think is challenging, is exciting, then you jump in. Of course, in the meantime, if you originally have a very good background in certain things, you may also like to keep going. Sometimes you make a complete swap, but sometimes you may swap back, but by doing something new, I think it will always give you some more credit and also more capability.

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

Oh, of course, something that got you excited, you feel is interesting even no matter whether other people got the same excitement or not? For example, in the early days, when I was doing deductive database, I got this, like a recursion, a very irregular recursion compilation, and I can solve, for example, the N-Queen problem in a rather declarative way, and I got very excited about that. But of course, you know, there are lots of new research topics you can work on. Like the first piece of work we did, this attribute-oriented induction. I also got pretty excited, at least at the time. Then the later ones… I probably believe we got this pattern growth method to solve like frequent patterns, sequential patterns, graph patterns, I believe those are pretty exciting things. And also those things we can see from the citation.

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

You mean outside of the research, altogether, or something related to research?

It could be either way.

Actually, if you say out of research, I actually quite like to travel, or climbing mountains, those things when I was young, I was quite energetic on those things.

Oh, I see, so when you said research or not research, I was imagining by not research you meant like you would volunteer to be the chair of SIGKDD or something, but you meant really outside the research, outside of work entirely! So you climb mountains!

Yes! But on the other hand, for research, I think you always try to find something exciting to work on. Quite often, I like to read for example, Scientific American. I always think there are a lot of different research topics you would love to know, and also you may try it. For example, in a lot of research, sometimes ideas you get are from those readings. You feel "oh, why this biologist can do this, why can't I do something similar?". I think a lot of research or knowledge can crossbreed.

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

Actually, in my early days, of course during the Cultural Revolution, it really broke out my dream. I actually would like to be a physicist, but I've never been able to get a chance. But I think it is interesting to read those things. But that is a different thing, I think once I got into computer science, I really love it.

Great! Thank you very much for talking with me today.

Thank you very much!