An Interview with Dr. Charu Aggarwal, SIGKDD Innovation Award Winner

ABSTRACT:

Dr. Charu Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York. He earned a Bachelor of Technology in Computer Science from the Indian Institute of Technology at Kanpur in 1993 and his Ph.D. in Operations Research, with a focus on mathematical optimization, from the Massachusetts Institute of Technology in 1996. He has worked extensively in the field of data mining, with particular interest in data streams, privacy, uncertain data and social network analysis. He has authored eight books, over 400 papers in refereed venues, and has applied for, or been granted, over 80 patents. Dr. Aggarwal sat down with *SIGKDD Explorations* to discuss how he first became involved in the KDD conference, his work at the IBM T.J. Watson Center and what excites him about the future of machine learning, data science and artificial intelligence.

CONGRATULATIONS ON RECEIVING THE SIGKDD INNOVATION AWARD! TELL US ABOUT YOURSELF. WHO IS DR. CHARU AGGARWAL?

I'm honored to be a distinguished research staff member at the IBM T.J. Watson Research Center, where my work focuses on machine learning. I did my undergraduate in computer science. Since my Ph.D. focused on optimization, probability and statistics, data mining seemed like a very natural career opportunity for me. So I started working in data mining at IBM when I finished my doctorate.

When I started working in 1996, the field was relatively new; KDD was only a year old at the time. Recently, I've been doing work on the machine learning and the deep learning side. Broadly, I work in industries, but I publish quite a lot, so I'm always looking for opportunities to transfer the research I do into practical applications. That is part of IBM's mission.

YOU HAVE AN H-INDEX OF 100. EXPERTS SAY AFTER 20 YEARS OF RESEARCH, AN H-INDEX OF 20 IS GOOD, 40 IS OUTSTANDING AND 60 IS TRULY EXCEPTIONAL. WHAT ADVICE WOULD YOU OFFER TO YOUNG ACADEMICS OR PROFESSIONALS SEEKING TO BE AS PRODUCTIVE AND IMPACTFUL IN THEIR PUBLISHED WORK?

My advice is to do research in a field you enjoy. If you do research in a field you like, it won't feel like work. It's very important that research not feel like a burden. Secondly, many conferences, like KDD, are quite selective in terms of the number of papers accepted. For young people, that can be discouraging. It's important to keep this in perspective and to recognize that acceptance rates may be low. Particularly during the early stages of your career, it's essential not to become discouraged. Broadly, if you are researching something interesting - something that inspires you - then publication and recognition will come naturally with time.

WAS IT DIFFICULT FOR YOU TO GET PUBLISHED WHEN YOU FIRST STARTED?

I had my share of papers rejected and that was fine. I accepted the fact that sometimes papers are selected and something they are not. The point is that publication should not be the primary goal. There has to be an interest in the process as a whole, then it is very hard to get discouraged.

BASED ON YOUR EXPERIENCE, WHAT TRENDS DO YOU FIND MOST INTERESTING?

Over the past twenty years, I've noticed some key trends with regard to the amount of data, both the growing amount available and techniques to deploy when there is a paucity of data. When I started working in 1996, the data sets were very small; maybe you would get a set of a few thousand points. These days, the data sets are simply huge. Streaming data is another aspect of the big data paradigm that began sometime in the early 2000s, around the time I started working in stream data mining.

All this led to a greater emphasis on deep learning, which inherently relies on massive volumes of data. Being able to train a neural network, for example, requires a substantial amount of data. I believe the trend towards deep learning is owed, in part, to superior computational power and greater availability of data.

The technology itself hasn't changed significantly. If you look at the architectural data networks, they aren't fundamentally different from what they were previously. While there have been small refinements, they are significantly deeper, as a result of now being able to train neural networks with many of layers. Progress over the past 10 to 20 years is linked to the availability of GPUs, which offer immense computational power and vast amounts of data.

One interesting aspect of this trend toward "bigger" data and greater computer power is even though big data has been a dominant theme, people are also talking about small data now. The greatest strength of a neural network is that it works very well with a lot of data, but this is also its weakness. If you're really going to simulate biology, you don't need a lot of examples. For example, a child recognizes a banana without a lot of examples. Neural networks don't seem to be able to do that, leading to another trend that is coming up in deep learning: how to predict outcomes or decisions with small data sets. Interestingly, we are kind of going backwards. We're asking, can we do one-shot learning or zero-shot learning? Are we able to train with very few examples?

YOU HAVE SERVED AS THE CO-CHAIR OF IEEE BIG DATA CONFERENCE, THE ICDM CONFERENCE, THE ACM CIKM CONFERENCE AND THE KDD CONFERENCE. IN YOUR OPINION, WHAT MAKES KDD UNIQUE?

Among the conferences that I have co-chaired, KDD is by far the largest, both in terms of the number of submissions and the number of attendees. It's a very popular conference because of the caliber of research presented. Papers published at KDD are often well cited and have significant impact long term.

Each year, KDD awards the "Test of Time" to the best paper from the past 10 years, and these papers have typically been cited thousands of times. Many of them are quite simply pioneering work. In that sense, KDD more than other conferences has papers that truly impact the industry. That's very impressive considering KDD is relatively new, compared to some of the other conferences I've worked on.

WHEN DID YOU FIRST DISCOVER KDD? WHAT ATTRACTED YOU TO THE COMMUNITY?

I focused on the theoretical when I did my Ph.D. At the time, when I was working at IBM, it made more sense for me to work in an applied domain. Given my background in optimization, probability and statistics, data mining and machine learning were a natural fit. In fact, I looked at some of the work being published and was immediately very interested because it was right in line with what I had been doing in my Ph.D. I was quickly drawn to the field.

Once I had published sufficiently, I was invited to join committees and eventually I was invited to be co-chair. I was always happy to help because if you are going to publish at a conference, obviously a lot of people are doing work for you – they are reviewing, editing and performing other tasks on your behalf. If someone invites you to do the work at the other end, then I believe it's something you have to do in order to give back. You have to do your share of the work.

WHAT IS YOUR FAVORITE PART OF THE KDD CONFERENCE?

The tutorials have always been my favorite part because I actually learn the most from these sessions. When you listen to a talk, which is usually around 20 minutes, you can get an idea of what they are doing, but at the end of the day, you still have to read the paper to actually learn what goes on. Tutorials, however, are extremely informative. You can learn a lot just by attending, especially the hands-on ones. KDD has one of the strongest tutorial programs of its kind, in my opinion.

WHAT EXCITES YOU MOST ABOUT THE FUTURE OF DATA SCIENCE?

Throughout my tenure in the field, I've continued to see surprising developments. When I started, being able to recognize an image in computer vision was considered a huge challenge. Today, you have neural networks that can recognize or classify an image better than a human can. This is not something I had dreamed of.

For example, the way chess programs were constructed twenty years ago, we never would have thought neural network approaches, like reinforcement learning, could actually perform better than a human. Now you have chess programs that play more like a human; the program takes risks, it makes sacrifices – the style of play is just like a human. The program even learns from experience, just like a human.

Given the unexpected innovation I've seen during my career, I'm fueled by anticipation for what the future holds. There's more I don't know about what might happen than what I do know, and that drives my excitement for what lies ahead.

WHAT DO YOU BELIEVE ARE THE BIGGEST CHALLENGES FACING THE INDUSTRY CURRENTLY?

One of the challenges I'm most interested in is the issue of big data versus small data. How do you learn when you don't have enough data? There is clearly work being done in this area; I've noticed more papers recently published on one-shot learning or zero spectrum learning.

In the early years, there were systems based on deductive reasoning and then inductive learning became more popular. Now there is an effort to create systems that integrate the best of both, which is exciting. If you look at how human beings do things, we reason with logic and we learn from examples; any machine that is going to be intelligent would have to have to be able to do both as well.

I'm also intrigued by the challenge of making learning more interpretable. For example, when a neural network makes a prediction, why did it make that decision? To a large extent, neural networks today are not easily interpretable, and they have been criticized for this. I have noticed a greater focus on this in the papers being published at KDD and look forward to future research on the topic.

IF YOU WERE TO GIVE THE NEXT GENERATION SOME ADVICE, WHAT WOULD IT BE?

I often see young attendees focusing primarily on what is trending at the moment. My advice is to not just select things on the basis of what is popular, but rather on the basis of what you find interesting. There is still work to be done in traditional machine learning or data mining. Generally, working on what you find most interesting will bring out the best in you.

Finally, don't be afraid to recommend yourself for a KDD committee. These days, the committees are always looking for people to get involved and participation is a great opportunity to collaborate with others in the field while also getting "behind the scenes" perspective on the inner working of planning this impressive annual conference.