Letter from the Impact Award Winner

Higher, Faster, Stronger: A Research Journey

Earlier this year, I received an email from the TCDE Awards Committee Chair informing me that I was the recipient of the 2021 IEEE TCDE Impact Award, with the citation "for contributions to deductive databases, streaming algorithms, and data integration." My first thought upon receiving this email was "Wow, this is a great honor!" My second thought was "Who on earth remembers my contributions to deductive databases?"

To put my cited contributions into the context of my research journey over the past three decades, and explain my use of "Higher, Faster, Stronger"[1] in the title of this letter, I summarize my cited contributions below.

Higher: My research in Deductive Databases had the goal of *elevating* query functionality in databases.

Faster: My research on Streaming Algorithms focused on *quickly* analyzing data streams in one pass.

Stronger: My research on Data Integration sought to make data combined from multiple web sources *robust*.

Deductive Databases: Ph.D. Studies I joined the University of Wisconsin–Madison to do my Ph.D. studies in 1987 after completing my B.Tech. in Computer Science & Engineering from IIT Bombay, India. Raghu Ramakrishnan had just joined the faculty there and he offered me the opportunity to work with him as a Research Assistant on a new deductive database system, CORAL, which he had started to build.

Deductive databases seek to enhance relational databases by adding the power of recursion to a relational query language with selection, projection, join, union and aggregation. This elevated declarative query language allows easy expression of a variety of natural problems, including shortest paths in graphs and bill of materials to assemble complex products. This enhanced query language expressiveness calls for time- and space-efficient query processing strategies, Along with Raghu and my fellow Ph.D. student, S. Sudarshan, we proposed and implemented many efficient and innovative strategies in CORAL. My Ph.D. thesis on "Representing and Querying Complex Information in the CORAL Deductive Database System" emerged from this body of work. I would like to thank Raghu for giving me the intellectual freedom to work on a research topic of my choice and Sudarshan for the many collaborations during our Ph.D. studies.

Since its heydays in the early 1990s, recursive query languages used in deductive databases have seen a resurgence in recent years, with diverse uses ranging from data integration and information extraction to networking, program analysis and security. It is intellectually satisfying to see research that I had done find new and interesting uses 2-3 decades later!

Streaming Algorithms: Processing Network Data at Scale After completing my Ph.D. from the University of Wisconsin–Madison in 1993, I joined AT&T Bell Labs in Murray Hill, New Jersey as a research scientist. When AT&T Bell Labs split up into AT&T Labs and Bell Labs in 1996, my affiliation changed to AT&T Labs, but my email address stayed the same, possibly offering a solution to the metaphysical identity problem of the Ship of Theseus. A few years later, I became the Head of Database Research at AT&T, with the good fortune of managing and working with a team of outstanding researchers.

Around the year 2000, many of us at AT&T were inspired by the idea of monitoring and analyzing streams of IP network packet headers using SQL-style declarative query languages, even though the high volume and velocity of the data precluded the possibility of persistently storing all the data flowing through AT&T's IP network. This idea inspired many of us to do a significant body of algorithms and systems research in the nascent area of data stream management systems over the next two decades, and it continues to be an active research topic even today. Since the data could not be persisted, it had to be processed and analyzed quickly in a single pass, using a

[1] Adapted from the Olympic Games motto, *Citius - Altius - Fortius*.

limited amount of memory. A variety of natural queries (such as finding quantiles of the distribution of round-trip times in the IP network) could be answered with approximation guarantees, but not exactly in the data stream model. The key researchers working on this topic were Graham Cormode, Lukasz Golab, Theodore Johnson, Flip Korn, S. Muthukrishnan, Vladislav Shkapenyuk and Oliver Spatscheck. We designed and implemented a large variety of streaming algorithms for fundamental query primitives, which were included in the GS data stream management system developed at AT&T and used to monitor network traffic. I would like to thank all of them for the fruitful collaborative research that we have engaged in over the years and decades.

Our ability to deploy streaming algorithms in a live system and tune their performance to operate at scale on AT&T's network data streams was crucial to make sure that these algorithms were both theoretically elegant and practically useful. These algorithms are now widely used to process data in a single pass in databases, networking, finance, e-commerce, and other domains.

Data Integration: Dealing with Web-Scale Heterogeneity Around the time that I joined AT&T Bell Labs in 1993, the World Wide Web was taking off – the first website was published in 1991 and the first web browser was released in 1992. With websites being built and independently populated with heterogeneous content, the challenge of building global information systems that could integrate data from multiple web sources and deal with web-scale heterogeneity inspired a new generation of researchers.

Web-scale heterogeneity comes in many flavors, ranging from schematic heterogeneity (where different web sources use different schemas to represent information) to syntactic data heterogeneity (where the same data values might be represented in different ways across web sources, e.g., typographical errors) to semantic data heterogeneity (where the web sources might even disagree on the correct values of data items). Robustly dealing with this range of web-scale heterogeneity has kept the data integration community busy for 25+ years now. Again, I have been extremely fortunate to work with very smart researchers on this challenging topic over the years. With Alon Halevy, we did some early work on resolving schematic heterogeneity, using the idea of "Local as View" mappings and answering queries using views in virtual data integration. With Marios Hadjieleftheriou, H.V. Jagadish and Nick Koudas, we developed efficient algorithms and tools for effectively dealing with syntactic data heterogeneity in data integration. With Laure Berti-Équille and Xin Luna Dong, we have conducted foundational research on resolving semantic data heterogeneity to perform truth discovery. I would like to thank all of them for the enjoyable and productive research collaborations over the last 25 years.

This body of data integration research has had considerable impact in academia and industry. As one example, the work on resolving semantic data heterogeneity to perform truth discovery influenced research on knowledge fusion and identifying trustworthy web sources at Google.

Research Journey: What is Next? Much of my current research is focused around data technologies for responsible data science and engineering. As society increasingly relies on data-driven decisions, we need to make sure that this decision making is trustworthy. Ongoing research ranges from timely, responsible data collection and data sharing, robust data curation for data to be fit for use, and responsible, transparent data use. There are enough important challenges here to keep me busy for a while!

In conclusion, I would like to thank those who nominated and endorsed me for this award as well as the TCDE Awards Committee. But most of all, I would like to thank my many collaborators over the years for making this research journey exciting and productive, and AT&T for giving me the opportunity and freedom to engage in curiosity-driven research!

Divesh Srivastava
AT&T, USA