Letter from the 2017 IEEE TCDE Impact Award Winner

This year I was honored to be given the Impact Award "For expanding the reach of data engineering within scientific disciplines." My interest in scientific applications started in the late 1980's when I met Dr. Chris Overton, who held a PhD in Developmental Biology and came to the University of Pennsylvania to complete a Master's in Computer and Information Science, because he believed that *the future of biology was computational* – quite a visionary for the time! After Chris was hired to head up the informatics component of the Center for Human Chromosome 22 in the 1990s, we frequently discussed the challenges he faced. This became a rich vein of research problems that the Database Group at Penn has worked on for over two decades. Most importantly, *addressing these challenges involved a close collaboration between end-users, systems builders and database experts.* Two of my favorite problems were data integration and provenance.

Data integration. Most data integration systems in the 1990's were built around relational databases, however genomic data was frequently stored in specialized file formats with programmatic interfaces. This led experts to state in a report of the 1993 Invitational DOE Workshop on Genome Informatics that "Until a fully relationalized sequence database is available, none of the queries in this appendix can be answered." However, the real problem was twofold: 1) integrating non-relational data sources; and 2) knowing what information was available and where. We answered the "unanswerable queries" within about a month using our data integration system, Kleisli, which used a complex-object model of data, language based on a comprehension syntax, and optimizations that went beyond relational systems. Our team also included experts who knew where the appropriate data sources were and how to use them. Since then, the database community has made excellent contributions in developing data integration systems that go well beyond the relational model; less progress has been made on knowing how to find the appropriate data sources and how to extract the right information.

Data provenance. Our team originally recognized the need for provenance when constructing an integrated dataset of genomic information: Not all data sources were equally trusted, but no-one wanted to express this opinion by failing to include a relevant data set. The solution was to make provenance available so that users could form their own conclusions. Since then, the importance of provenance has been widely recognized, especially as it relates to reproducibility and debugging, and the database community has made excellent progress in "coarse-grained" provenance for workflows, "fine-grained" database style provenance, and "event-log" style provenance. However, the usability of provenance remains a challenge: provenance is collected more often than it is used!

Bioinformatics is just a precursor of the "tsunami" that is now Data Science, and many even more interesting challenges lie ahead – see the CRA report on "Computing Research and the Emerging Field of Data Science" (available at http://cra.org). As before, these problems are best addressed by teams of people working together. I am encouraged to see our community rising to these challenges, and expanding the chain of end-users, systems builders and database experts to include statisticians and machine learning researchers, among many other types of experts required in developing solutions to real problems in Data Science.

Susan B. Davidson
University of Pennsylvania