Letter from the Rising Star Award Winner

It is an honor to receive the TCDE Rising Star Award "for contributions to interactive data-intensive systems for exploratory data analysis." Thank you to the awards committee as well as my nominator, letter writers, mentors, collaborators, graduate and undergraduate advisees, and of course my family for supporting my career. I could not have done this without you!

Designing systems for exploratory data analysis requires knowledge of how to efficiently clean, process, and model data at scale, as well as a deep understanding of how people interpret and manipulate this data to extract insights and make decisions. In this letter, I highlight four major challenges at the intersection of data management, visualization, human-computer interaction, and machine learning that I believe will shape the future of exploratory data analysis research.

How Can We Model Users More Efficiently? With the constant stream of advancements in artificial intelligence and machine learning, a natural question is: how can we leverage existing modeling techniques to anticipate and provision for users' interactions within data exploration systems? A few projects explore the use of decision trees, Markov chains, and Markov decision processes (including my own work), but alternative techniques such as neural network architectures are difficult to apply to abstract sequences of user interaction logs. A major hurdle in this space is that we struggle to collect enough data to sufficiently train large interaction models. How can we collect enough data to support current modeling techniques, or how can we design models to use minimal interaction data in behavior-driven optimization contexts?

How Can We Increase the Transfer of Innovations Across Areas? My research often involves translating innovations in visualization and HCI, such as theories of human analysis behavior and principles of interface design, into practical data processing optimizations of interest to the data management community. However, much of this work has to be done manually, because we lack effective processes for translating ideas across these areas. I see opportunities to cross pollinate not just research ideas, but entire methodologies between communities. For example, tight integration between theory and systems work has led to significant breakthroughs in database research. In a similar fashion, developing *systematic processes* for translating theory in visualization and HCI into hints to database management systems is an exciting research direction. One approach could be developing *programmable* theories of visualization and interaction behavior, which could then be compiled to query languages such as SQL.

How Can We Design More Modular and Integrative Tools? Data scientists are constantly changing their workflows in response to shifts in data analysis environments, modalities, and best practices. As a result, there is no one size fits all solution to data science, and the "winners" tend to be the tools that offer the greatest flexibility, one example being Jupyter Lab. I am happy to see our community shift towards supporting computational notebook environments such as Jupyter Lab, but to maximize our impact, we need to shift our design methodology away from monolithic systems towards *modular* libraries and *integrative* toolkits. With this approach, data scientists can pick and choose which components to use and easily connect them together as needed. Unfortunately, this is the opposite of how many visualization systems and database management systems are designed, requiring a paradigm shift across areas.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

How Can We Design More Socially Responsible Tools? Our community works hard to make data science tools faster, more efficient, more robust, and more fair. We should be proud of our achievements, but we still have a long way to go. For example, our innovations lead to tangible improvements in the daily work of data scientists and data engineers, but what about the tools *they* build using our technology, and the people impacted by them? How do we design data science tools that maximize societal benefits while minimizing potential societal harms? And how can we develop pathways within the data science pipeline for those affected to promote data equity and seek data justice?

I hope this letter provides a new perspective on human-centered data science work, and encourages you to take action towards building a better world. My work has widened the path towards this goal, as recognized by this award, but it takes a community to execute paradigm shifts and make lasting change. And we can't do this alone. By partnering with other critical areas such as human-computer interaction, visualization, and machine learning, we can build a positive legacy in data science.

Leilani Battle
University of Washington, USA