Letter from the Rising Star Award Winner

I am delighted and honored to receive this award "for designing and deploying data analytics systems powered by innovative machine learning and artificial intelligence algorithms." I am thankful to my nominator, letter writers, and the award committee, as well as my students, mentors, and colleagues who helped me in my work. This letter lays out my research worldview at the exciting intersection of DB+ML systems. My hope is to rouse more interest in this intersection, which is fast growing in importance within the wider computing landscape.

Past Waves of DB+ML Systems Work. The DB world has studied ML systems issues for over 20 years. I have seen, and contributed to, three waves. The first wave (late 1990s to late 2000s) saw the rise of *in-RDBMS data mining tools*. Their main focus was to scale ML algorithms to DB-resident data without modifying RDBMS internal code. A key example is Oracle Data Mining. The second wave (late 2000s to mid 2010s) saw the rise of *unified ML system templates*. Their main focus was to simplify ML implementation on RDBMSs and dataflow systems. Key examples are MADlib, the Bismarck system from Wisconsin, and Spark MLlib.

What is New Now? The third wave, from mid 2010s, is a much bigger tidal wave. What changed? First, ML and AI have now become a ubiquitous business-critical need, not some arcane academic curiosity. Second, "Big Data" and cloud computing have expanded the variety, complexity, and scale of DB+ML problems. Third, deep learning has revolutionized ML itself, unlocking unstructured data for analytics. And last but not least, this third wave is characterized by "thinking outside the DBMS box" to study problems in *contexts that matter for more ML users*, viz., bringing DB ideas to ML systems and applications, not just bringing ML to DBMSs.

The New DBfication of ML/AI. I have interacted with over three dozen data/ML practitioners across enterprise companies, Web companies, and domain sciences, as well as many DBMS and cloud vendors, both to help deploy my research to practice and to learn about new problems and bottlenecks. In my opinion, this third wave is not a one off but a historic tectonic realignment in computing: *ML/AI is being fundamentally re-imagined as DB-style workloads*. This requires new science to more deeply understand the phenomena, processes, and tradeoffs involved, as well as new technology rooted in that science to raise ML user productivity, reduce resource needs and costs, enable new applications, ensure compliance with laws and societal values, etc. The closest analogy is how the "RDBMS research community" formed around the relational model and SQL c. 1980s.

DB problems exist across the *entire end-to-end lifecycle of ML/AI* applications. Based on my conversations I split that lifecycle into three main stages: *Sourcing* of data for ML, *Building* of prediction functions, and *Deployment* of such prediction functions. Let me give a few key examples of DB problems in each stage. In the Sourcing stage, we need new tools for easier data acquisition for ML on data lakes, less manual data preparation tools for ML, and higher throughput data labeling tools. In the Building stage, we need new model+data debugging schemes and query optimization techniques to improve resource efficiency of ML systems. In the Deploy stage, we need less manual monitoring of predictions and orchestration of complex data+ML pipelines.

Becoming a DBesque ML/AI "Savior." The above problems are not "pure ML" problems or "pure DB" problems' but "DB+ML problems." But ML folks cannot tackle them on their own without painfully reinventing ideas already familiar to DB folks, e.g., metadata management, query optimization, provenance, etc. If the DB world fails to help the ML/AI world, is it not a shirking of our intellectual responsibility to the computing field, to science itself, and to broader society? Of course, other computing communities (e.g., systems, HCI, and PL) are also studying some of these problems. But the DB community has a unique expansive role to play due to two major reasons: (1) it has the most *successful track record in commodification* of data-centric software, and (2) it is a *vertical slice of all of computing* to study data-centric software in an eclectic and holistic way spanning systems, theory, algorithmics, empiricism, hardware, interfaces, applications, and benchmarking.

In conclusion, I hope this letter gave you a new perspective on DB+ML systems work. My own work, recognized by this award, is but one part of this tidal wave. It will take a whole research community to unlock the full potential of data-centric computing for our modern data-driven world. I hope the DB community can rise to this occasion, reach out with humility to the ML/AI world to learn and engage deeply with ML/AI applications, build solid systems, work with practitioners, and translate our research to impact on practice.

Arun Kumar
UCSD, USA