

## Letter from the Editor-in-Chief

The last time the Data Engineering Bulletin published a special issue on the topic of Knowledge Bases was more than six years ago. In the September 2016 issue, we asked leading researchers in this field for their thoughts on the present and future of Knowledge Bases.

At the time, there were two major focuses in the Knowledge Base field: data harvesting and heterogeneous data management. The primary concern for data harvesting was how to automatically build large, high-quality knowledge bases from Internet sources. The most difficult challenge in heterogeneous data management was determining the best way to model the data in a knowledge graph that allows applications to easily access knowledge.

It is clear that these two challenges are still pressing today. Building knowledge graphs remains a task with a low ROI because the universe of knowledge is too vast, and many methods are as impractical as boiling the ocean. Even though information extraction techniques have matured with the progress of machine learning in text processing, we still struggle to capture the tail facts, or facts mentioned only once or twice in the corpus. Furthermore, the issue of heterogeneous data management continues to stymie applications' ability to leverage knowledge graphs.

Nonetheless, much progress has been made in this area since our September 2016 issue. There has been a strong pull from applications due to their increasing needs for information and knowledge, and simultaneously a strong push from technology, fueled by the great process in machine learning, particularly in natural language processing.

The Data Engineering Bulletin will devote several issues to review the field of Knowledge Bases. This latest issue, curated by Yangqiu Song, shares a glimpse of the knowledge base landscape today. We are witnessing a shift of focus and methodology, with representation, data quality, and reasoning taking precedence. The first paper in the issue, for example, focuses on data incompleteness, which is a critical problem for large-scale knowledge graphs. Given the limitations of knowledge harvesting, the approach of combining facts and ML-powered estimations for every information need is promising. The second paper addresses another recurring issue of knowledge graphs: data inconsistency. It allows users to compare different outcomes of reasoning as the result of the inconsistency.

Haixun Wang  
Instacart