CrowdDB: Query Processing with People and Machines

Michael Franklin, UC Berkeley

The challenge of "Big Data" analytics is not just data size -- it's about issues such as diversity, ambiguity, and incompleteness in both queries and the underlying data.

While advances in scalable data processing are helping to address the data size problem, there remain many important data-centric tasks where humans are more proficient than current state-of-the art algorithms. Crowdsourcing has emerged as a major problem-solving and data-gathering paradigm that provides the ability to leverage human intelligence and activity at large scale. Emerging popular crowdsourcing platforms have programmatic interfaces (APIs) that provide the opportunity to create hybrid human/computer systems for data-intensive applications. An ongoing effort to better understand the development of such hybrid computation systems, the CrowdDB project uses human input via crowdsourcing to process queries that neither database systems nor search engines can adequately answer. While CrowdDB leverages many aspects of traditional database systems, there are also important differences from both an implementation and conceptual perspective. In this talk, I'll present some recent results on CrowdDB (built with colleagues at ETH Zurich and developed as part of the U.C. Berkeley AMPLab) and developing hybrid human/machine query processing systems. I'll also share an overview of the broader AMPLab research agenda, which is focused on building a data analysis infrastructure that seamlessly integrates Algorithms, Machines and People.