ABSTRACT: Today, we collect huge amount of data. Video sharing websites, retailing corporations, airline industry are just a few examples. The volume of the data we collect grows faster than the growth of the computational power, and the size of data is projected to increase faster in the future. Performing data analytics over terabytes or petabytes of data, distributed over hundreds of machines, has become a norm. We obtain valuable information and insights by analyzing data, and those analyses are important bases for our decision making.

The growth in data size, however, inevitably increases query latencies. Horizontal scaling alone is not sufficient for achieving real-time data analytics, especially when the size of data grows faster than the computational power. Approximate query processing (AQP) intends to produce query answers in real-time at the cost of small quality losses in query answers. AQP is useful when we prefer obtaining approximate answers (e.g., with 1% error) within a few seconds compared to obtaining exact answers in hours.

In my defense, I will show we can greatly speed up this AQP by learning from past computations and data. Specifically, my PhD work enhances three types of AQP—aggregation, searching, and visualization—by exploiting past computations and by building task-aware data synopses. For exploiting past computations and building task-aware synopses, my work incorporates statistical inference and optimizations techniques into the data analytics systems. The contributions in my work resulted in up to 20x speedups for many data analytics tasks.

Chairs: Profs. Michael Cafarella and Barzan Mozafari