

K-means for Large Data *Chris Mueller, September 9, 2005*

K-means is the most widely used clustering algorithm. The increasing availability of large databases and streaming and distributed data has driven the development of novel k-means implementations. Two important classes are single pass and parallel algorithms. Single pass algorithms (ideally) process each data item only once and are useful for streaming data and very large data sets. Parallel algorithms take advantage of multiple processors, local or remote. A SQL implementation is also included to highlight the use of declarative languages and industry-centric applications of k-means.

