The Web, social media, mobile activity, sensors, Internet commerce and so on all provide many extremely large datasets from which information can be gleaned by data mining. This book focuses on practical algorithms that have been used to solve key problems in data mining and can be used on even the largest datasets.

It begins with a discussion of the MapReduce framework, an important tool for parallelizing algorithms automatically. The tricks of locality-sensitive hashing are explained. This body of knowledge, which deserves to be more widely known, is essential when seeking similar objects in a very large collection without having to compare each pair of objects. Stream processing algorithms for mining data that arrives too fast for exhaustive processing are also explained. The PageRank idea and related tricks for organizing the Web are covered next. Other chapters cover the problems of finding frequent itemsets and clustering, each from the point of view that the data is too large to fit in main memory, and two applications: recommendation systems and Web advertising, each vital in e-commerce.

This second edition includes new and extended coverage on social networks, machine learning and dimensionality reduction. Written by leading authorities in database and web technologies, it is essential reading for students and practitioners alike.