In the new era of “big data,” we are increasingly faced with the challenges of processing vast volumes of data. Given the limits of individual machines (compute power, memory, bandwidth), increasingly the solution is to process the data in parallel on many machines. This course focuses on the fundamentals of the “data-parallel” approach to scaling computation to handle common data analytics tasks. You will learn about the challenges of parallelism, communication, synchronization, and coordination; programming models for performing certain kinds of computation in a scalable way across many compute nodes; common approaches to converting algorithms to such programming models; and popular distributed frameworks for analytics tasks such as filtering, graph analysis, clustering, and classification. Prerequisites: prior programming experience (CIS 110, CIT 590, or equivalent). Basic familiarity with data structures, graphs, graph traversals, probability.