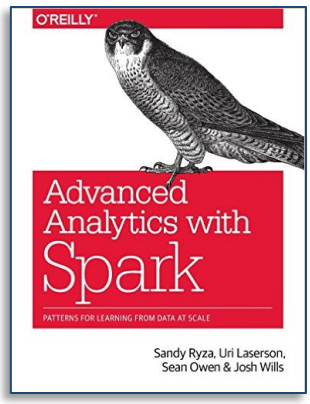
CSE 398/498 Big Data Analytics

Spring 2017 • Tu 2:35 – 3:50 (2 credits) • Packard Lab 216
Professor Dan Lopresti • Computer Science and Engineering

In this 2-credit project course, we will gain a practical working knowledge of large-scale data analysis using the popular open source Apache Spark framework. Spark provides a powerful model for distributing programs across clusters of machines and elegantly supports patterns that are commonly employed in big data analytics, including classification, collaborative filtering, and anomaly detection, among others.

Working from the course textbook, we will study and program solutions for problems including: music recommender systems; predicting forest cover with decision trees; anomaly detection in network traffic with K-means clustering; understanding Wikipedia with Latent Semantic Analysis; analyzing co-occurrence networks with GraphX; geospatial and temporal data analysis on the New York City Taxi Trips data; estimating financial risk through Monte Carlo simulation; analyzing genomics data and the BDG project; and analyzing neuroimaging data with PySpark and Thunder.

Supplemental readings will provide additional background for each application area, but most of the work in the course will involve implementing, studying, and enhancing the programming examples from the textbook. During class, students will take turns presenting their own solutions and helping to lead the discussion. A final project will be required.



Enrollment in this course is limited and requires permission of the instructor. Please note that this is not a basic course on data mining, cluster computing, or programming in Scala; it assumes you already know something about these topics and/or you can learn them quickly on your own. Contact the instructor, Professor Dan Lopresti, for details.

2-CREDIT COURSE FOR SPRING '17