CSE 398/498 BIG DATA ANALYTICS

Spring 2017 • 2:35 pm – 3:50 pm Tu • Packard Lab 216

Instructor **Professor Daniel Lopresti**
Email dal9@lehigh.edu ~ Ext 85782
Office Hours 2:00 pm – 4:00 pm on Wednesdays (or by appointment) in PL 350

Texts *Advanced Analytics with Spark: Patterns for Learning from Data at Scale* by Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills, O'Reilly Media, 2015, ISBN-13: 978-1491912768. (primary)

Programming in Scala by Martin Odersky, Lex Spoon, and Bill Venners, Artima Inc., 2011, ISBN-13: 978-0981531649. (if needed)

CourseSite Materials and discussion forums will be available @ http://coursesite.lehigh.edu/

Grading

• Homework assignments	50 points (25%)
• Class participation	50 points (25%)
• Final project, presentation, and write-up	100 points (50%)

Notes Homework assignments are generally due at 12:00 midnight on the Monday night before our class. All work must be submitted on time to receive full credit by uploading a .zip archive to CourseSite containing the programs you have written and your notes for the week. The late penalty is -10% per day or fraction thereof. The final project write-up will be 5 pages for students enrolled in CSE 398, and 8 pages for students enrolled in CSE 498.

Accommodations for Students with Disabilities If you have a disability for which you are or may be requesting accommodations, please contact both your instructor and the Office of Academic Support Services, Williams Hall, Suite 301 (610-758-4152) as early as possible in the semester. You must have documentation from the Academic Support Services office before accommodations can be granted.

Principles of Our Equitable Community Lehigh University endorses The Principles of Our Equitable Community.* We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.

Academic Integrity The work you submit in CSE 398/498 must be entirely your own. While I encourage you to discuss basic concepts with friends and classmates, plagiarism is never acceptable. Such cases will be referred to the University Committee on Discipline and, if you are found guilty, you may be given the failing grade WF in the course. If you have questions about this policy at any point, ask me. It is far better to be safe than sorry when your academic career may be on the line.

* http://www.lehigh.edu/~inprv/initiatives/PrinciplesEquity_Sheet_v2_032212.pdf

CSE 398/498 BIG DATA ANALYTICS

Spring 2017 • 2:35 pm – 3:50 pm Tu • Packard Lab 216

Course Schedule

Date	Topic	Readings*	Note
Jan. 24	Course Intro		
Jan. 31	Hands-on Intro to Scala and Spark	Chapters 1, 2	
Feb. 7	Music Recommender Systems Using Least Squares	Chapter 3	
Feb. 14	Predicting Forest Cover with Decision Trees	Chapter 4	
Feb. 21	Anomaly Detection in Network Traffic with K-means Clustering	Chapter 5	
Feb. 28	Understanding Wikipedia with Latent Semantic Analysis	Chapter 6	
Mar. 7	Analyzing Co-occurrence Networks with GraphX	Chapter 7	
Mar. 14	No class		Spring Break
Mar. 21	Geospatial and Temporal Data Analysis on the NYC Taxi Trip Data	Chapter 8	Possible date switch
Mar. 28	Estimating Financial Risk through Monte Carol Simulation	Chapter 9	
Apr. 4	Analyzing Genomics Data and the BDG Project	Chapter 10	
Apr. 11	Analyzing Neuroimaging Data with PySpark and Thunder	Chapter 11	
Apr. 18	Final project presentations I		
Apr. 25	Final project presentations II		
May 2	Course Wrap-Up		Possible date switch

* All readings are taken from *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. In addition to reading each chapter carefully along with the associated supplemental materials I will assign throughout the course, you must work through the programming examples in the book on your own. After you have programmed what you find in the book, you must implement your own extensions, enhancements, and variations which you will be asked to explain and demonstrate for the class. Creativity will be rewarded! Participation during classes and contributing to the discussions will constitute a significant portion of your grade.