The University of Georgia

Department of Computer Science

Course Information Sheet

CSCI 4360

Data Science II

Brief Course Description (50-words or less)	This course introduces the students to advanced analytics techniques in data science, including random forests, semi-supervised learning, spectral analytics, randomized algorithms, and just-in-time compilers. Students are also introduced to distributed and out-of-core processing.
Extended Course Description / Comments	This course aims to provide students with deep knowledge of sophisticated data science techniques for making sense of data across domains. Students are instructed how to process data that is incomplete or missing, use hybrid techniques to analyze such as semi-supervised learning, and are introduced to distributed programming using Hadoop and Spark. Furthermore, students are given the opportunity to explore just-in-time compilation, both in Python and in the new scientific computing language Julia. The course is appropriate both for students preparing for research in Data Mining and Machine Learning, as well as Bioinformatics, Science and Engineering students who want to apply Data Mining techniques to solve problems in their fields of study.
Pre-Requisites and/or Co-Requisites	CSCI 2360 Data Science I
Required, Elective or Selected Elective	Selected Elective Course
Approved Textbooks (if more than one listed, the textbook used is up to the instructor's discretion)	Author(s): Richert, Willi and Luis Pedro Coelho Title: Building Machine Learning Systems in Python Edition: 1st Edition, 2013 ISBN-13: 978-1782161400 Author(s): Jake VanderPlas Title: Python Data Science Handbook Edition: 1st Edition, 2016 [expected] ISBN-13: 978-1491912058
Specific Learning Outcomes (Performance Indicators)	This course builds on the concepts from Data Science I by introducing students to more advanced analytics techniques. At the end of the semester, all students will be able to do the following: 1. Design and implement a full data science pipeline, from data preprocessing and feature selection to model evaluation and performance optimization. 2. Rigorously and quantitatively select the optimal model for a given problem. 3. Move between Python and Julia to employ the strengths of each. 4. Select existing packages or employ techniques to handle analysis of data that is too large to load into memory at once. 5. Scale analyses beyond single cores to highly parallel and fully distributed heterogeneous computing environments.

Relationship Between Student Outcomes and Learning Outcomes

		Student Outcomes										
		a	b	c	d	e	f	g	h	i	j	k
Learning Outcomes		●	●							●	●	●
		●	●	●							●	
			●							●	●	
		●	●							●		●
		●	●							●	●	●

Major Topics Covered
(Approximate Course Hours)

3 credit hours = 37.5 contact hours
4 credit hours = 50 contact hours

Note: Exams count as a major topic covered

Introduction and statistics review (7.5-hours)

Information theory (2.5-hours)

Decision trees and random forests (5-hours)

Collecting, formatting, and integrating data (2.5-hours)

Structured vs unstructured data (2.5-hours)

Randomized algorithms (5-hours)

Semi-supervised learning and label propagation (2.5-hours)

Spectral analytics (7.5-hours)

Out-of-core data processing (2.5-hours)

Just-in-time compilation and Julia (5-hours)

Introduction to Hadoop and Spark (7.5-hours)

Course Master

Dr. Shannon Quinn