The University of Georgia

Department of Computer Science

Course Information Sheet

CSCI 3360

Data Science I

Brief Course Description (50-words or less)	This course will be a rigorous overview of methods for text mining, image processing, and scientific computing. Core concepts in supervised and unsupervised analytics, dimensionality reduction, and data visualization will be explored in depth.
Extended Course Description / Comments	Provides students with an in-depth dive into data science with the Python language. Students will learn the theoretical justifications for statistical techniques in classification, clustering, and dimensionality reduction, learn how to implement them within the Python framework, and apply them to extracting knowledge from scientific datasets. Upon completion of this course, students will be able to design an algorithmic pipeline, leverage existing Python packages and incorporate their own to implement the pipeline, and visualize the results of their computations.
Pre-Requisites and/or Co-Requisites	(CSCI 1301 Introduction to Computing and Programming OR CSCI 1360 Foundations of Informatics and Analytics) AND (MATH 2250 Calculus I for Science and Engineering OR CSCI 2150 Introduction to Computational Science)
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): Joel Grus Title: Data Science from Scratch: First Principles with Python Edition: First Ed., 2015 ISBN-13: 978-1491901427 Author(s): Willi Richert and Luis Pedro Coelho Title: Building Machine Learning Systems with Python Edition: 1st ed., 2013 ISBN-13: 978-1782161400
Specific Learning Outcomes (Performance Indicators)	This course provides students with the core tools for extracting knowledge from data. At the end of the semester, all students will be able to do the following: 1. Use existing Python tools to read and preprocess raw data of various formats (text, images, binary). 2. Choose the optimal statistical model for extracting knowledge from a particular dataset, given the advantages and disadvantages of the model. 3. Implement at least one algorithm from the categories of regression, classification, clustering, and convex optimization. 4. Design and document analytical pipelines to be reproducible by others. 5. Use and interpret the results of dimensionality reduction on high-dimensional datasets.

6. Choose the most effective visualization to convey the knowledge learned from the data.

Relationship Between Student Outcomes and Learning Outcomes

		Student Outcomes										
		a	b	c	d	e	f	g	h	i	j	k
Learning Outcomes	[illegible]	●								●		
	[illegible]	●	●								●	
	[illegible]	●		●							●	●
	[illegible]	●	●			●					●	●
	[illegible]	●									●	
	6	●					●			●		

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 to data science (2.5-hours)

Preprocessing text and image data (2.5-hours)

Regression (7.5-hours)

Supervised classification (7.5-hours)

Unsupervised clustering (7.5-hours)

Dimensionality reduction (7.5-hours)

Convex optimization (5-hours)

Data visualization (5-hours)

Introduction to “big data” (2.5-hours)

Notebooks for scientific reproducibility (2.5-hours)

Course Master

Dr. Jaewoo Lee