School of Computing
UNIVERSITY OF GEORGIA

Course Information Sheet

CSCI 4360/6360

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

Use this section to put additional information that's relevant to whom this course is targeting

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 3360 Data Science I

Approved Textbooks
(If more than one, course text used during a semester is at the discretion of the instructor)

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)

These are a (non-exhaustive) list of specific, measurable outcomes, as they relate to the course & program objectives.

These learning outcomes should avoid using ambiguous language such as "understand" or "familiar".

Performance indicators must include an action verb (indentifying the depth to which students should demonstrate performance), and the content referent that is the focus of the instruction (from ABET)

Target number 5 - 10

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. Understand the technical, ethical, and logistical trade-offs of some models over others for certain situations.
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 Course Outcomes and Learning Outcomes

	ABET Learning Outcomes						
Specific Learning Outcomes		A	B	C	D	E	F
	1	●	●			●	●
	2	●	●				
	3		●	●	●	●	
	4	●	●				●
	5	●	●				●

Program Outcomes

(These are ABET-specified and should not be changed)

1. Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.
2. Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline.
3. Communicate effectively in a variety of professional contexts.
4. Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.
5. Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.
6. Apply computer science theory and software development fundamentals to produce computing-based solutions.

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)
Deep learning (12-hours)
Generative models (5-hours)

Assessment Plan for this Course

Each time this course is offered, the class is initially informed of the Course Outcomes listed in this document, and they are included in the syllabus. At the end of the semester, an anonymous survey is administered to the class where each student is asked to rate how well the outcome was achieved. The choices provided use a 5-point Likert scale containing the following options: Strongly agree, Agree, Neither agree or disagree, disagree, and strongly disagree. The results of the anonymous survey are tabulated and results returned to the instructor of the course.

The course instructor takes the results of the survey, combined with sample student responses to homework and final exam questions corresponding to course outcomes, and reports these results to the

	ABET committee. If necessary, the instructor also writes a recommendation to the ABET committee for better achieving the course outcomes the next time the course is offered.
How Data is Used to Assess Program Outcomes	Each course Learning Outcome, listed above, directly supports one or more of the Program Outcomes, as is listed in "Relationships between Learning Outcomes and Program Outcomes". For CSCI 4360/6360, Program Outcomes (1), (2), (3), (4), (5), and (6) are supported.
Course Master	Dr. Shannon Quinn
Course History	Last modifies 1/30/2024 By Dr. Shannon Quinn