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

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

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

Author(s): Jake VanderPlas
Title: Python Data Science Handbook

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 (identifying 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. 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 Course Outcomes and Learning Outcomes

<table>
<thead>
<tr>
<th>Program Outcomes</th>
<th>Learning Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>•</td>
</tr>
<tr>
<td>2</td>
<td>•</td>
</tr>
<tr>
<td>3</td>
<td>•</td>
</tr>
<tr>
<td>4</td>
<td>•</td>
</tr>
<tr>
<td>5</td>
<td>•</td>
</tr>
</tbody>
</table>

## Program Outcomes

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

- **a.** An ability to apply knowledge of computing and mathematics appropriate to the discipline.
- **b.** An ability to analyze a problem, and identify and define the computing requirements appropriate to its solution.
- **c.** An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs.
- **d.** An ability to function effectively on teams to accomplish a common goal.
- **e.** An understanding of professional, ethical, legal, security and social issues and responsibilities.
- **f.** An ability to communicate effectively with a range of audiences.
- **g.** An ability to analyze the local and global impact of computing on individuals, organizations, and society.
- **h.** Recognition of the need for and an ability to engage in continuing professional development.
- **i.** An ability to use current techniques, skills, and tools necessary for computing practice.
- **j.** An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modeling and design of computer-based systems in a way that demonstrates comprehension of the tradeoffs involved in design choices.
- **k.** An ability to apply design and development principles in the construction of software systems of varying complexity.

## Major Topics Covered

(Approximate Course Hours)

<table>
<thead>
<tr>
<th></th>
<th>3 credit hours = 37.5 contact hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction and statistics review</td>
<td>(7.5-hours)</td>
</tr>
<tr>
<td>Information theory</td>
<td>(2.5-hours)</td>
</tr>
<tr>
<td>Decision trees and random forests</td>
<td>(5-hours)</td>
</tr>
<tr>
<td>Collecting, formatting, and integrating data</td>
<td>(2.5-hours)</td>
</tr>
<tr>
<td>Structured vs unstructured data</td>
<td>(2.5-hours)</td>
</tr>
<tr>
<td>Randomized algorithms</td>
<td>(5-hours)</td>
</tr>
<tr>
<td>Semi-supervised learning and label propagation</td>
<td>(2.5-hours)</td>
</tr>
<tr>
<td>Spectral analytics</td>
<td>(7.5-hours)</td>
</tr>
<tr>
<td>Out-of-core data processing</td>
<td>(2.5-hours)</td>
</tr>
<tr>
<td>Just-in-time compilation and Julia</td>
<td>(5-hours)</td>
</tr>
<tr>
<td>Introduction to Hadoop and Spark</td>
<td>(7.5-hours)</td>
</tr>
</tbody>
</table>

Note: Exams count as a major topic covered.
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 (a), (b), (c), (i), (j), and (k) are supported.

Dr. Shannon Quinn

Course Master
Course History