LUX ET SPES

STONEHILL COLLEGE
LEO J. MEEHAN SCHOOL OF BUSINESS

Fall 2020 – DAN 606
Statistics for Data Analytics (4 cr.)

Faculty Information:

Course Instructor:	Dr. Robert Harbert, Assistant Professor of Biology & Bioinformatics
Telephone:	Cell: 540-354-8104
Office:	Shields Science Center 204
Virtual Office Hours (Zoom):	Zoom appts. can be made via email or at https://calendly.com/rharbert/scheduling **Request evening appointments by email. I am happy to make arrangements outside of M-F 9-5**
Email:	rharbert@stonehill.edu
Course Slack Workspace:	https://stonehillmpsda21.slack.com
Course Meeting Time/Place:	Fridays 5-9pm on scheduled residency weekends in Meehan 103

Leo J. Meehan School of Business Mission Statement:

Anchored in the tradition of the Congregation of Holy Cross, the Leo J. Meehan School of Business offers a high-quality business education. Strong communication skills, business literacy, and a global perspective are emphasized to prepare students to make ethical, thoughtful, and significant contributions to their organizations and communities as professionals. Building on the foundation of a liberal arts education, and using engaged mentorship, we provide a multidisciplinary curriculum delivered with an emphasis on the student.

Vision:

The Leo J. Meehan School of Business seeks to be a leader in business education emphasizing an experiential learning environment and ability to graduate professionals who can contribute and lead with purpose in a rapidly changing global business environment.

Core Values:

The Leo J. Meehan School of Business fosters an inclusive community which honors the dignity of all persons consistent with the mission of Stonehill College and in the spirit of the Congregation of Holy Cross. The School…

1. Delivers a high-quality education that incorporates best practices.
2. Builds on the foundation of a liberal arts education.
3. Promotes a high degree of engaged mentorship.
4. Incorporates global and ethical perspectives into student learning.
5. Encourages quality intellectual contributions and professional activities that advance the teaching/pedagogy, theory, and practice of business.
6. Pursues knowledge creation through collaboration with other disciplines.

Competencies:

The following competencies guide the delivery of our Master of Professional Studies in Data Analytics program:

- *lg1* *Identify a business problem or opportunity and how data analytics may be applied to solve the problem and/or increase business value.*
- *lg2* *Acquire, access, assay, and prepare data for analysis.*
- *lg3* *Identify and perform appropriate methods of data analysis.*
- *lg4* *Interpret and communicate analysis results to stakeholders without bias.*
- *lg5* *Conduct data analysis with high regard for security, privacy, and ethics.*

Course Description/Objectives:

Covers the impact of big data on business and what insights big data can provide through hands-on experience with the tools and systems used by big data scientists and engineers. Software basics in Hadoop and Spark (with discussion of related software). By following along with provided code, students will experience how one can manage analytics and predictive modeling for large data sets. By the end of the course students will be able to perform basic big data analysis on a large provided data set.

At a high level, the main learning objectives/topic areas of this course are:

Topic coverage may change at the discretion of the instructor.

- Basic interaction with the Unix (and related) command line systems
- Hardware and software solutions
- Setting up and managing a Hadoop distributed file system
- Interacting with Hadoop via Pig/Spark/Etc.
- Performing predictive modeling with Spark
- SAS integration with Hadoop

Responsibilities:

- bring a willingness to learn and contribute to the overall atmosphere of collegiality in this course;
- keeping up with the **readings** as indicated in the course schedule;
- attending all **weekend residencies**;
- participating in **online activities and discussion**;
- other assignments as assigned

Prerequisites:

- None

Student Resources:

Required Resources

Textbooks, Documentation, and White Papers:

- **Textbook:** "*Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale"*
  Authors: Mendelvitch, Stella, and Eadline
  Publisher: Addison-Wesley Professional
  ISBN: 978-0134024141

Software:

.

- Access to the Boston University Shared Computing Cluster (SCC) provided through collaboration with the Massachusetts Green High Performance Computing Center

  Account setup will be covered in class

- Microsoft Office

Please **do not** use Google Docs, Apple Pages, or any other office suite. If you do not have the latest Microsoft Office installed, please see the instructions located at: https://stonehill.teamdynamix.com/TDClient/1841/Portal/KB/ArticleDet?ID=73312. If you have any problems downloading or installing Microsoft Office, please email the IT Service Desk at service-desk@stonehill.edu or call 508.565.1111.

Online Resources:

- SAS.com

https://SAS.com provides access to e-Learning courses, software documentation, videos, discussion communities, and other helpful resources. You should get to know the SAS.com website and use it to access the latest content from SAS. All students should have a SAS.com profile.

- eLearn

eLearn will allow you to post completed assignments as well as view homework, quiz, and exam grades. eLearn can be found at: http://elearn.stonehill.edu. Login with your Stonehill College username and password. You can also access eLearn from your myHill account.

SAS Certifications:

- SAS Certified Specialist: Base Programming Using SAS 9.4

This course will help prepare you for portions of the *SAS Certified Specialist: Base Programming Using SAS 9.4* certification credential. All students in the program will be required to sit for a SAS certification exam of their choosing during the summer semester. More information on the SAS Global Certification Program can be found at https://www.sas.com/en_us/certification.html or on our Stonehill SAS Community Group site at https://studentsstonehill.sharepoint.com/sites/sasGroup/

Do not pay for any certification courses, materials, or exams on SAS.com. Your enrollment in the program grants you access to all of these resources at no cost.

******* Late Assignment *******

- **All assignments are due by the date and time listed on the course schedule in eLearn.** All work is submitted online via an eLearn dropbox unless stated otherwise in the schedule or in class.
- You are responsible for submitting all assignments on time even if you are absent from class.
- **Late work will be accepted for a limited time with a 20% per day penalty.**
- **NO late work will be accepted 5 days past the listed due date.**

Grade Determination:

Assignments and homework (30%)

- In-class/homework problems assigned with each module.

Online Discussion (15%)

- Semi-weekly class discussion on #dan606 slack channel.

Exams (15%)

- Two practical exams given online at the midpoint (after residency XX) and at the end of the course. The second exam will not be cumulative.

Code Portfolio (30%)

- You will keep and document a portfolio of all of your code files for in-class and homework assignments in a repository on https://github.com.

Community Contribution (10%)

- Participation during residencies as well as online forums in constructive and professional ways counts towards successful completion of this course.

Credit will be determined by assigning a numerical value to each category, corresponding to 100%. Final grades will be calculated by multiplying the relative weights by the achievement earned for each category. A letter grade will be assigned, using the following table:

Achievement	Letter Grade	Definition	Quality Points per Credit Hour
95-100	A	Excellent, work that is of the highest standard, showing distinction and meets acceptable standard for graduation	4.00
90-94	A-		3.70
87-89	B+	Good, work that is of high quality and meets acceptable standard for graduation	3.30
83-86	B		3.00
80-82	B-	Satisfactory, work that fulfills requirements in quality and quantity and meets acceptable standard for graduation	2.70
77-79	C+	Unsatisfactory, for *required* or *core* coursework and does not meet acceptable standard for graduation Acceptable, only for one *elective* course and meets acceptable standard for graduation	2.30
73-76	C	Unsatisfactory, work that does not fulfill requirements or meet acceptable standard for graduation, and considered failing grade for required graduate coursework	2.00
70-72	C-		1.67
67-69	D+		1.33
60-66	D		1.00
<60	F	Failure, work undeserving of credit	0.00

Honor Code & Academic Integrity:
In the context of community of scholarship and faith and anchored in the belief in the inherent dignity of each person, the students, faculty, staff and administration of Stonehill College maintain an uncompromising commitment to academic integrity. We promote a climate of intellectual and ethical integrity and vigorously uphold the fundamental values of honesty, trust, fairness, and responsibility while fostering an atmosphere of mutual respect within and beyond the classroom. Any violation of these basic values threatens the integrity of the educational process, the development of ideas, and the unrestricted exchange of knowledge. Therefore, we will not participate in or tolerate academic dishonesty.

Violations of the academic integrity policy include but are not limited to the following actions:

- Presenting another's work as if it were one's own;
- Failing to acknowledge or document a source even if the action is unintended (i.e., plagiarism);
- Giving or receiving, or attempting to give or receive, unauthorized assistance or information in an assignment or examination;
- Using cheating websites disguised as "homework or study help" sites. Downloading from or uploading to these sites is a violation of this policy;
- Fabricating data;
- Submitting the same assignment in two or more courses without prior permission of the respective instructors;
- Having another person write a paper or sit for an examination;
- Unauthorized use of electronic devices to complete work;
- Furnishing false information, including fabricating excuses for incomplete work or lying about violating this policy; or
- Not reporting someone who you know has violated this policy.

I take academic honor and honesty very seriously. If you are unsure as to whether your actions in this course are in accordance with the Stonehill College Honor Code, please ask me before carrying out these actions. Violation of these policies may result in sanctions up to and including failure of the course or expulsion.

For further information on this policy, the procedure for adjudicating incidents, and your rights as a student, see: http://catalog.stonehill.edu/content.php?catoid=9&navoid=405&returnto=search#stonehill_college_academic_honor_code_policy_procedures

Accommodations:
Stonehill College is committed to providing a welcoming, supportive and inclusive environment for students with disabilities. The Office of Accessibility Resources (OAR) provides a point of coordination, resources and support for students with disabilities and the campus community. If you anticipate or experience physical or academic barriers based on disability, please let me know so that we can discuss options. You are also welcome to contact OAR to begin this conversation or to establish reasonable accommodations for this or other courses. OAR is located within the Academic Services & Advising Suite in Duffy 104. For additional information please call 508.565.1306 or email accessibility-resources@stonehill.edu. ***If you plan on using accommodations in this class, please see me at the beginning of the semester. It is your responsibility to communicate with me so that we can plan the use of your accommodations.***

Cell Phone Policy:

Please be courteous to your faculty and fellow students and avoid using your cell phone during class.

Netiquette:

This course is taught in a hybrid manner. A hybrid course is one in which some course instruction and activities take place in the face-to-face classroom (instructor and students together in one location) and some take place online. When you are participating in the online portion of this course, I expect that you follow the rules of Netiquette. You can view these rules here.

Tentative Course Schedule:

** Full schedule of readings and assignments will be posted on course eLearn at least one week before each residency.

Date	Agenda	To Do
1/22	Orientation – Introductions & Course Policies *Module 1: What is "Big Data" anyways? Or "I know it when I see it"*	*Goal(s)* • Defining (and recognizing) Big Data • Understanding hardware and software solutions • Welcome to the Unix command line (tips and tricks for handling data) *Read* • Chapter 1-2
2/12	*Module 2: Distributed File Systems*	*Goal(s)* • Explore the Apache big data software ecosystem (Hadoop, Pig, Spark, Hive, Sqoop, Flume, etc.) • Set up Hadoop File System • Import data into Hadoop cluster • Explore data in Hadoop *Read* • Chapter 3-4
3/5	*Module 3: Handling Data and Optimizing Hadoop File Systems*	*Goal(s)* • Deal with data quality issues • File system organization and optimization • Basic data analysis with HDFS *Read* • Chapter 5
3/26	*Module 4: SAS Proc Hadoop*	*Goal(s)* • Integrate Hadoop with SAS *Read* • *SAS Proc Hadoop Documentation* • *https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3405-2019.pdf*
4/16	*Module 5: Predictive Modeling*	*Goal(s)* • Implement predictive models and ML with Spark *Read* • Chapter 7-8
5/7	*Module 6: Clustering and Anomaly Detection*	*Goal(s)* • Implement basic clustering algorithms • Implement anomaly detection *Read* • Chapter 9-10