Introduction to Data Science
Data Science 101
Fall 2013

Professor: Paul Anderson, Ph.D.
Office: 212 J.C. Long
Office Hours: T: 2 – 3 PM
R: 2 – 3 PM
W: 12 – 1 PM
And open door policy

Contact Information: andersonpe2@cofc.edu
Office Phone: 953-8151
Facebook: andersonpe2@cofc.edu
Facebook group: https://www.facebook.com/groups/324697457664695/

Course Description: Introduction to the use of computer based tools for the analysis of large data sets for the purpose of knowledge discovery. Students will learn to understand the Data Science process and the difference between deductive hypothesis-driven and inductive data-driven modeling. Students will have hands-on experience with various on-line analytical processing and data mining software and complete a project using real data.

Required Text: Data Mining with R: Learning with Case Studies by Luis Torgo.

Course Website and Schedule: http://birg.cs.cofc.edu/index.php/Fall_2013_DATA_101

Course (learning) outcomes:
1. To gain an overview the field of knowledge discovery
2. To be able to distinguish and translate between data, information, and knowledge
3. To learn how to store, query, aggregate data in databases
4. To be able to distinguish problems based on computability
5. To learn how to implement distributed computing and storage
6. To apply algorithms for inductive and deductive reasoning
7. To learn introductory and state-of-the-art data mining algorithms
8. To apply data mining, statistical inference, and machine learning algorithms to a variety of datasets, including text, image, biological, and health
9. To apply information filtering on real world datasets
10. To apply information validation on real world datasets
11. To apply artificial intelligence concepts to real world datasets
12. To understand the social, ethical, and legal issues of informatics and data science

Course Prerequisite: The course has no prerequisites other than your interest and IT skill fluency. The course is accessible to all students who have an interest in a breadth-first overview of the field of discovery informatics.

Course Web Page: All materials associated with this course will be posted on the course web page, Dropbox, or through OAKS.

Homework Policy: Homework will be assigned on Thursday and due the following Thursday. There will be no late homework assignments as they will build upon each other, and solutions will be posted immediately.

Exam Policy: Student performance will be assessed through one midterm and one comprehensive final exam, as indicated on the course schedule. Both will be a combination of in class written and take home.

Late Assignment Policy: Everyone will be granted one missed lab assignment; however, you will be required to complete all homework assignments by their originally scheduled due date. Failure to do so without appropriate documentation will
result in a zero on that assignment. No exceptions will be made because solutions will be posted immediately following grading.

**Grading Policy:**

1. Midterm 30%
2. Homework 30%
3. Comprehensive Final 30%
4. Lab 5%
5. Kaggle 5%

**Semester Data Science Project:** Pairs of students will be required to participate in weekly standup meetings. This will be a graded exercise. Each pair will select a competition from the website [www.kaggle.com](http://www.kaggle.com). They will be required to lead a discussion on their competition for 5 – 10 minutes every Thursday. These discussions are expected to build upon previous weeks with additional material being brought in to delve deeper into the topic. While you are expected to lead your segment of this meeting, you will also be graded on your participation during the time of other groups. Students who successfully submit their own solution to one of the competition will be awarded a bonus on the final exam.

**Grading Scale:** A: 90-100; B: 80-89; C: 70-79; D: 65-69; F: <65. Plusses and minuses will be used at the discretion of the instructor.

**Cheating:** Students are expected to work independently in this course. Collaborations on specific assignment details are a violation of the honor code. Use of another student's answers is considered cheating, and cases of this nature will be taken to the Judicial Board.

**Attendance Policy:** Attendance at all lectures is required. Excused absences for illness, personal/family emergency or academic/professional commitments will be granted at the discretion of instructor.

**Disability Accommodation:** Any student who feels that he or she may need an accommodation due to a disability should speak to me individually to discuss your specific needs. For additional help please contact the College of Charleston Center for Disability services at [http://www.cofc.edu/~cds/](http://www.cofc.edu/~cds/).

**Electronics Devices:** The use of electronic devices, both stand-alone and network capable, will play an increasingly important roll in teaching and learning at the College of Charleston, including their use in our classrooms. The following policy specifies which electronic devices and network connections can be used and when their use is disallowed in this class.

Devices that are **allowed** to be used at certain times:

*During class, except during tests and exams:*
Allowed are mobile computing devices, e.g. laptops, palmtops, tablets, electronic pens, calculators. Mute the sound. The use of these devices is encouraged for accessing WebCT, taking notes and running simulations during class.

*During tests, exams and quizzes:*
No electronic devices are allowed to be powered up, unless otherwise specified by the instructor. All books and notes are to be stowed below desk level.

**Network Access:** Students may use wired, WiFi and IR networks available during class, whenever electronic devices are allowed, provided the use of the network does not distract other students or the instructor.

Be considerate and sensitive to others. All student behaviors are subject to the policies in the College of Charleston Student Handbook.
Course Manifesto

I teach my courses under a simple philosophy of two parts: (i) always put the student first, and (ii) continually learn as much as I can about the subject.

Every decision I make is based to increase your knowledge on the subject. This is often obvious, but at other times, it is less so.

You can always ask me to explain my philosophy or reasoning behind anything related to the course.

You can always discuss the grading of an assignment, in a professional manner of course.

I am a casual professor who is chronically unorganized and constantly excited about trying new ways to improve your course. Some students respond really well to this blend of excitement and resultant chaos. Some do not. If there is something that could improve your individual learning without adversely affecting others in the course, I will do my best to accommodate it. I always want to hear about it.

You may be wondering why I feel it is necessary to tell you all of this. It is because I can’t resist trying to innovate each semester. This leads to a little chaotic at times, but it will also hopefully lead us to greatness. Or at least sometime better than ordinary.

This course is designed to give you a real taste of data science. We will be presented a survey and history of the field, followed by an application driven curriculum. For this purpose, we will be focusing on techniques related to those presented in the book Data Mining with R: Learning with Case Studies. This will be supplemented with related material. Each week will basically have two halves. The first half will consist of a traditional lecture style. The second half will be for stand-up discussions on Kaggle projects, lab time, and introduction to this week’s homework. You will work the lab time in pairs, which I will assign based on background. We will also be using a new experimental website developed by my research group called Learn2Mine. This being the first deployment of the software, expect some delays and technical glitches.

Enjoy!!!