Syllabus: CAP4770/5771, Fall 2016
Introduction to Data Science

Catalog Description: See UF course catalog.
Credits: 3 **Grading Scheme:** Letter
Introducing the basics of data science including programming for data analytics, file management, relational databases, classification, clustering and regression. The foundation is laid for big data applications ranging from social networks to medical and business informatics.

Pre-requisites and Co-requisites:

Data Structures and Algorithms (COP3530) or equivalent.

Optional but preferred: Information and Database Systems I (CIS 4301) and an introductory course to probabilities and statistics are pre-requisites.

Course Objectives:

In order to address the growing need from both industry and academia (e.g., medical and bio informatics, financial, law enforcement, economics, decision support, social networks) for big data analytic skills including, data management, data mining, machine learning and data visualization, this is the first of the three-course series in the Data Science curriculum. The aim is to bring students, with basic programming and data structure background, to be abreast with common tools used for Data Science application development. This course serves as an introduction to the basics of data science including programming for data analytics, file management, relational databases, classification, clustering and regression. The foundation is laid for big data applications ranging from social networks to medical and business informatics.

Instructor:

Prof. Daisy Zhe Wang, Office: E456, Phone: (352) 505-7626, E-mail: daisyw@cise.ufl.edu
Office Hours: Mon/Wed 4-5pm or by appointment, Location: CSE456/457.

Course Information:

- **Credits:** 3
- **Section:** 228F/037B
- **Meeting Times:** MWF: 8th (3:00 to 3:50pm)
- **Where:** CSE Building, E119
- **Teaching Assistant:** Xiaofeng Zhou, Dihong Gong
- **Laboratory:** N/A
- **Material and Supply Fees:** None
- **Class web page:** https://ufl.instructure.com/courses/331578

TA1 Xiaofeng Zhou
Contact: TBA, Office Hour: TBA, Location: TBA

TA2 Dihong Gong
Contact: TBA, Office Hour: TBA, Location: TBA

Textbooks and Software Required: We will be using Amazon Web Services (AWS) and software supported on top of AWS. AWS credits will be available to each student.

Recommended Readings from:

a) Data Science from Scratch (DSS), Joel Grus, O'Reilly Media Inc., http://shop.oreilly.com/product/0636920033400.do

b) Python for Data Analysis (PDA), Wes McKinney, O'Reilly Media Inc., http://proquest.safaribooksonline.com/9781449323592

c) Mining of massive datasets (MMD), A. Rajaraman and J.D. Ullman, Cambridge University Press, 2011. ISBN-10: 1107015359, ISBN-13: 978-1107015357, http://www.mmds.org/ (public online access)

d) Natural Language Processing with Python (NLTK Book): http://www.nltk.org/book/ (public online access)

Further Reading:

e) Learning scikit-learn: Machine Learning in Python (MLP), Guillermo Moncecchi and Raul Garreta,

f) Doing Data Science (DDS), Cathy O'Neil and Rachel Schutt, O'Reilly Media Inc., http://proquest.safaribooksonline.com/9781449363871

Course Outline and Topics:

This course will give an introduction to the basic data science techniques including programming in Python, SQL/SPARQL and Map-Reduce for small and big data manipulation and analytics. We also cover topics including data collection, data preparation, data querying, data analytics including pattern mining, classification, clustering, data visualization, and parallel computing platforms. We will also touch on advanced data analytics including NLP, knowledge extraction, graph analytics, graph querying, knowledge bases and crowd sourcing. During the course, we will also make an effort to introduce key application areas of data science including business intelligence, social media, biomedical informatics, computational ecology and e-discovery.

Course Schedule (tentative and subject to change)

Date	Topic of Lectures and Labs	Readings	Assignment
M 8/22	Class logistics and introduction	DSS Chap 1	
W 8/24	Lec 1: Data Science Overview	DSS Chap 2	Lab 0 Material
F 8/26	Lab 0: Python Programming Bootcamp	DSS Chap 9 PDA Chap 6	
M 8/29	Lec 2: Data types, collection and preparation		
W 8/31	Lec 2: Data types, collection and preparation		Lab 1 Material
F 9/2	Lab 1: Data Preparation (Unix)		
M 9/5	Holiday, no class	DSS Chap 3 & 7 PDA Chap 8	
W 9/7	Lec 3: Exploratory Data Analysis with Statistics and Visualization		
F 9/9	Lec 3: Exploratory Data Analysis with Statistics and Visualization		Lab 2 Material
M 9/12	Lab 2: Exploratory Data Analysis (Python, Pandas & matplotlib)	DSS Chap 10 PDA Chap 5 & 7	
W 9/14	Lec 4: Tabular Data Processing, Matching and Integration		
F 9/16	Lec 4: Tabular Data Processing, Matching and Integration		Lab 3 Material
M 9/19	Lab 3: Joining Multiple Tables (Python & Pandas)	DSS Chap 14 & 16 & 17	
W 9/21	Lec 5: Supervised Models: Classification, Regression		
F 9/23	Lec 5: Supervised Models: Classification, Regression		Lab 4 Material
M 9/26	Lab 4: Classification and Regression (Python & Scikit)	MMD Chap 2	
W 9/28	Lec 6: Scaling up analytics, map-reduce		
F 9/30	Lec 6: Scaling up analytics, map-reduce		Lab 5 Material
M 10/3	Lab 5: Map-Reduce (Python)		NIST DSE Guidelines
W 10/5	Lab 8: Getting NIST DSE data		Midterm Review Questions
F 10/7	Review		
M 10/10	midterm		
		NLTK Chap 3 & 6 & 7	
W 10/12	Lec 7: Extracting Information from Textual Data		
F 10/14	Lec 7: Extracting Information from Textual Data		Lab 6 Material
M 10/17	Lab 6: Information Extraction from Text (python/NLTK)	MMD Chap 5	
W 10/19	Lec 8: Graph Analysis and Random Walk?		
F 10/21	Lec 8: Graph Analysis and Random Walk		Lab 7 Material
M 10/24	Lab 7: Page Rank (Map Reduce)	DSS Chap 10 PDA Chap 5 & 7	
W 10/26	Lec 9: Outlier detection and Data Cleaning		NIST Cleaning Guidelines I
F 10/28	Lec 9: Outlier detection and Data Cleaning		
M 10/30	Lab 9: Cleaning Task I	DSS Chap 11, MLP Chap 1 & 2	
W 11/2	Lec 10: Maching Learning: Models and Processes/Big Data Science Systems		NIST Cleaning Guidelines II
F 11/4	Homecoming, no class		
M 11/7	Lab 10: Cleaning Task II		
W 11/9	Holiday, no class		
F 11/11	Lec 11: Feature Extraction, Engineering and Clustering	DSS Chap 19, MLP Chap 3 & 4	NIST Prediction Guidelines I
M 11/14	Lec 11: Feature Extraction, Engineering and Clustering		
W 11/16	Lab 11: Prediction Task I	DSS Chap 11, MLP Chap 1	
F 11/18	Lec 12: Machine Learning: Performance Metric and Evaluation		NIST Prediction Guidelines II
M 11/21	Lec 12: Machine Learning: Performance Metric and Evaluation		
W 11/23	Thanksgiving, no class		
F 11/25	Thanksgiving, no class		
M 11/28	Lab 12: Prediction Task II		NIST Guidelines III
W 11/30	Advanced Topics: Automatically Constructing Knowledge Bases		
F 12/2	Advanced Topics: Knowledge Base Applications and data science projects		
M 12/5	Final Lab: Improving Cleaning and Prediction III		
W 12/7	Conclusion and final discussions		
12/8 - 12/9	Reading days		
12/10 - 12/16	Exam week		Final Project Report Due

Attendance and Expectations:

- **We require class attendance and participation, since most of the class material will be delivered in class. Roll-calls are conducted in class/labs via canvas.**
- **Moreover, in-class lab assignments and homework will be conducted to test the understanding of the material via canvas.**
- **Personal laptop computers are required in class for participation in labs and assignments.**
- Please return your labs/homework/projects in time. Late returns will cause 20% deduction in your grade for that lab/homework/project for each late day.
- If I postpone or cancel the office hour, I will post it in the announcements.
- Please avoid any activities that will disturb the flow of the lectures: Silence your cell phones, pagers, etc.
- Excused absences are consistent with university policies in the undergraduate catalog (https://catalog.ufl.edu/ugrad/current/regulations/info/attendance.aspx) and require appropriate documentation.

Grading Policy – Methods of Evaluation

~7 In-class Labs & Homework (45%)
Midterm (20%)
Final Project (30%)
Attendance (5%)

Grading Scale:

Roughly the boundaries will be:

Percent	**Grade**	**Grade Points**
90.0 - 100.0	A	4.00
87.5 - 89.9	A-	3.67
85.0 - 87.4	B+	3.33
80.0 – 84.9	B	3.00
77.5 - 79.9	B-	2.67
75.0 - 77.4	C+	2.33
70.0 – 74.9	C	2.00
67.5 - 69.9	C-	1.67
65.0 - 67.4	D+	1.33
60.0 - 64.9	D	1.00
0 - 59.9	E	0.00

More information on UF grading policy may be found at:
https://catalog.ufl.edu/ugrad/current/regulations/info/grades.aspx

Policy on Missed Quizzes and Late Assignments

I am extremely sympathetic if you have some conflict that will make it difficult for you to attend a quiz or get an assignment completed on time. However, I don't like people to take advantage of my sympathy. So I have a very strict and explicit set of rules governing missed quizzes and late assignments. To be fair to everyone in the class, these rules are always followed to the letter and without exception, so don't even ask!

1. If you have some conflict and feel like you may need an extra day on the assignment or need to take the quiz a day or two later, it **must be cleared with me no fewer than one week (seven days) before the assignment is due or the quiz will be held**. I am generally sympathetic to the standard excuses, if you sound credible ("I have three exams that day", "My brother's bar mitzvah is the day before the assignment is due", are the usual type of excuses that I hear). However, if you do not clear it a week in advance, you will receive a zero, with only a two exceptions, given below.

2. If you are ill at the time of a quiz or right before an assignment is due and so you miss the one week window, or if there is a death in your immediate family, I will allow a late assignment or a make-up provided (a) you can give me **proof of the circumstances**, and (b) you let me know **before the quiz is held or the assignment is due**.

3. If you simply don't turn in the assignment or don't show up for the quiz, the only valid excuse is a note from a **doctor given as proof that you were injured or ill at the due date to such an extent that it**

would be unreasonable for you to send me an email or leave a message letting me know of your illness or injury. In any other case, the result is a zero on the assignment.

Policy on Regrade Requests:

Much of the grading in the class is subjective. As such, no grade is viewed as final when you first receive it. It is your responsibility to look over every paper that is returned and to carefully check to make sure that it was graded correctly. You are free to discuss grading orally with me or one of the TAs. However, **any request for an actual regrade must be made in writing within one week of the time that the paper is returned, with no exceptions**. **All regrade requests must be prepared using a word processing program; a hard copy should then be submitted to me or a TA, along with the original graded work**. On your regrade request, carefully describe why you feel that you were scored unfairly and/or incorrectly. Even if you discussed the grading issue orally with someone, **the written discussion must be self-contained and be able to evaluated based only on what is written on the paper**.

Honesty Policy:

UF students are bound by The Honor Pledge which states, "We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honor and integrity by abiding by the Honor Code. On all work submitted for credit by students at the University of Florida, the following pledge is either required or implied: "On my honor, I have neither given nor received unauthorized aid in doing this assignment." The Honor Code (https://www.dso.ufl.edu/sccr/process/student-conduct-honor-code/) specifies a number of behaviors that are in violation of this code and the possible sanctions. Furthermore, you are obligated to report any condition that facilitates academic misconduct to appropriate personnel. If you have any questions or concerns, please consult with the instructor or TAs in this class.

Software Use:

All faculty, staff and student of the University are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against University policies and rules, disciplinary action will be taken as appropriate. We, the members of the University of Florida community, pledge to uphold ourselves and our peers to the highest standards of honesty and integrity.

Students Requiring Accommodations

Students with disabilities requesting accommodations should first register with the Disability Resource Center (352-392-8565, https://www.dso.ufl.edu/drc) by providing appropriate documentation. Once registered, students will receive an accommodation letter which must be presented to the instructor when requesting accommodation. Students with disabilities should follow this procedure as early as possible in the semester.

Student Privacy

There are federal laws protecting your privacy with regards to grades earned in courses and on individual assignments. For more information, please see: http://registrar.ufl.edu/catalog0910/policies/regulationferpa.html

Course Evaluation

Students are expected to provide feedback on the quality of instruction in this course by completing online evaluations at https://evaluations.ufl.edu/evals. Evaluations are typically open during the last two or three weeks of the semester, but students will be given specific times when they are open. Summary results of these assessments are available to students at https://evaluations.ufl.edu/results/.

Campus Resources:

Health and Wellness

U Matter, We Care:
If you or a friend is in distress, please contact umatter@ufl.edu or 352 392-1575 so that a team member can reach out to the student.

Counseling and Wellness Center: http://www.counseling.ufl.edu/cwc, and 392-1575; and the University Police Department: 392-1111 or 9-1-1 for emergencies.

Sexual Assault Recovery Services (SARS)
Student Health Care Center, 392-1161.

University Police Department at 392-1111 (or 9-1-1 for emergencies), or http://www.police.ufl.edu/.

Academic Resources

E-learning technical suppor*t*, 352-392-4357 (select option 2) or e-mail to Learning-support@ufl.edu. https://lss.at.ufl.edu/help.shtml.

Career Resource Center, Reitz Union, 392-1601. Career assistance and counseling. https://www.crc.ufl.edu/.

Library Support, http://cms.uflib.ufl.edu/ask. Various ways to receive assistance with respect to using the libraries or finding resources.

Teaching Center, Broward Hall, 392-2010 or 392-6420. General study skills and tutoring. https://teachingcenter.ufl.edu/.

Writing Studio, 302 Tigert Hall*,* 846-1138. Help brainstorming, formatting, and writing papers. https://writing.ufl.edu/writing-studio/.

Student Complaints Campus*:* https://www.dso.ufl.edu/documents/UF_Complaints_policy.pdf.

On-Line Students Complaints*:* http://www.distance.ufl.edu/student-complaint-process.