CS 459: Data Science Programming

G. Exoo

Spring 2019

E-mail: ge@cs.indstate.edu
Office Hours: MWF 12:15-1:00, 4:00-4:15
Office: A-140A Root Hall
Final Exam: May 08, 3:00PM.

Web: **cs.indstate.edu/CS459**
Class Hours: MWF 3:00-3:50pm
Class Room: A-017 Root Hall

Course Description

Data Science is a relatively new discipline at the interface of Mathematics, Statistics and Computer Science. It has emerged because of the importance of massive datasets in social media, retailing, telecommunications, network and computer administration, and the sciences, especially Physics, Astronomy, and Biology. Data Science can be viewed as a five stage process:

- Data preparation
- Data exploration
- Data visualization
- Data modeling
- Model testing and evaluation

The course will deal with the tools that are used in each of these phases, and will and will focus on working with large datasets. It will also provide an introduction to the **R** programming language.

Learning Outcomes

Recommended Textbooks

R Programming for Data Science

by Roger D. Peng Publisher: Lean Publishing, 2015.

Learning R

by Richard Cotton Publisher: O'Reilly Media, 2013.

The Art of R Programming

by Norman Matloff Publisher: No Starch Press, 2011.

Hands-On Programming with R

by Garrett Grolemund Publisher: O'Reilly Media, 2014.

R for Data Science

by Hadley Wickham and Garrett Grolemund Publisher: O'Reilly Media, 2017

R in Action

by Robert I. Kabacoff Publisher: Manning Publications, 2011

R in a Nutshell

by Joseph Adler Publisher: O'Reilly Media, 2010

The Linux Command Line

by William E. Shotts, Jr. Publisher: No Starch Press, 2018.

Online Texts/References

There is a wealth of free material online. At the *R Documentation* site one can find links to freely available books and tutorials on R. Many of the books focus on applications of R to a particular area of study.

Linux/Unix for Beginners

R Documentation

Expected Amount of Work

If you take this class seriously and get what you should out of it, some weeks you will likely be spending an average of 15 hours/week or more on the class. The students who get A's in their CS courses and have an easy time finding jobs spend at least this much time. Not everyone will need to spend this much time and not all weeks will be the same, but you should plan on putting in whatever time it takes. Note - your classes should be more important than your part-time job.

Course Announcements

Announcements regarding the course will be made both during class, on the class web page, and to your cs email (which you should learn how to use). Communications related to grades will be sent to your @sycamores.indstate.edu email address. You should regularly check these locations.

Classroom Conduct

You may not use cell phones, iPods/music players, etc. during class. You should be civil and respectful to both the instructor and your classmates, and you should arrive to class a few minutes before the scheduled lecture so you are ready for lecture to begin on time. You may use your computer during class if you are using it to follow along with the examples that are being discussed. You may not check email, facebook, work on other courses, etc. during class.

Course Outline

The main topics in the course are listed below. This is intended as a topical outline, not a timeline. Database and Machine Learning concepts will be developed together. A dynamic timeline for the class can be found at the end of this document.

Course Topics

- Introduction to Data Science
  - Tools for Data preparation
    - The Linux shell
    - Regular expressions
    - *grep*, *sed*, *awk*, etc.
  - Data exploration
    - Review of elementary statistical concepts
    - Using R as a statistics application
  - Data Visualization
    - Fundamentals of computer graphics
    - Using R as a graphics environment
  - Data modeling
    - Predictive modeling
    - Data mining
    - Machine learning
- R Programming

  - The language
    - R data types
      - Vectors, sequences, lists, data frames, matrices.
    - R control structures
    - R functions
    - R input and output
    - Classes
  - Improving performance
    - Some basic concepts from Computer Science
      - Complexity, running time, and the big-O notation
    - Interface to other languages

- Project
  - Preparing the data
  - Understanding the
  - Modeling the data
  - Testing the model

Assignments

The students in this course have the following responsibilities: read assigned readings before lecture, attend lecture, complete homework assignments, take in-class quizzes, take exams, and complete all projects. In this class, your final grade will be based primarily on the projects, and the lesser extent on the exams and quizzes. CS Course Policies Note that this course follows all standard CS course policies. In particular check the CS course policies related to - cheating/plagiarism, attendance, missing exams. See http://cs.indstate.edu/info/policies.html for details.

All assignments are posted in a pdf file on the class web page. Each such file will indicate the number of points, the due date and time, and the location where your assignment should be saved. Failure to save your work in the correct location will be viewed as equivalent to not doing the work.

Late Assignments

The maximum points you will receive for late assignments will decay exponentially with time. If n is the number of points the assignment is worth, then a perfect assignment that is between $d-1$ and d days late will net at most $n/(2^d)$ points. We suggest attempting a homework assignment the day it is given, or the day after, so that if you do not understand how to do the assignment, you will have time to seek help. You may need to ask for help more than once, and you should certainly plan on spending a lot of time in the CS lab (A-015 Root Hall). Many of the homework assignments require thought and creative problem solving. Do not expect to solve the problems the first time you attempt them.

Grading Policy

We try to design homework assignments and exams so that a standard cutoff for grades will be close to what you deserve. After the first exam a grade will be created in Blackboard called **Letter Grade** that is intend to be your current grade in the class. The grades are generally based on the following table.

A	93-100
A-	90-93
B+	87-90
B	83-87
B-	80-83
C+	77-80
C	73-77
C-	70-73
D+	67-70
D	63-67
D-	60-63
F	0-60

Grades are intended to indicate your mastery of the course material. The following are offered as guidelines.

A

You can do all the assignments on your own.

B+/A-

The student understands almost everything, and should be able to use this knowledge in other courses or in a job.

B-/B

The student understands most, but not all, topics well.

C/C+

Learned enough and have the minimum skills to move on in the subject.

D+/C-

The student made some effort in, and understands some things at a high level, but hasn't mastered the details well enough to be able to use this knowledge in the future.

D-

Students will normally not get an F if - they attend 80% of the lectures, complete some of the assignments up through the end of the course, and get nearly half of the problems on the final exam correct.

F

Normally, students that get an F simply stopped doing the required work at some point.

Blackboard

The course has a blackboard site. You should see this course listed under your courses for the current term. The blackboard site is used only for giving you your grades (go to the course in blackboard, then click *My Tools*, and then *My Grades*.

Academic Integrity

Read CS course policies in terms of what is and is not allowed on assignments: http://cs.indstate.edu/info/policies.html. Please ask the instructor if you have doubts about what is considered cheating in this course.

Special Needs / Student Disabilities

Indiana State University recognizes that students with disabilities may have special needs that must be met to give them equal access to college programs and facilities. If you need course adaptations or accommodations because of a disability, please contact us as soon as possible in a confidential setting either after class or in my office. All conversations regarding your disability will be kept in strict confidence. Indiana State University's Student Support Services (SSS) office coordinates services for students with disabilities: documentation of a disability needs to be on file in that office before any accommodations can be provided. Student Support Services is located on the lower level of Normal Hall in the Center for Student Success and can be contacted at 812-237-2700, or you can visit the ISU website under A-Z, Disability Student Services and submit a Contact Form. Appointments to discuss accommodations with SSS staff members are encouraged. Once a faculty member is notified by Student Support Services that a student is qualified to receive academic accommodations, a faculty member is obligated to provide or allow a reasonable classroom accommodation under ADA.

Disclosures Regarding Sexual Misconduct

Indiana State University fosters a campus free of sexual misconduct including sexual harassment, sexual violence, intimate partner violence, and stalking and/or any form of sex or gender discrimination. If you disclose a potential violation of the sexual misconduct policy I will

need to notify the Title IX Coordinator. Students who have experienced sexual misconduct are encouraged to contact confidential resources listed below. To make a report or the Title IX Coordinator, visit the Equal Opportunity and Title IX website: http://www.indstate.edu/equalopportunity-titleix/titleix.

The ISU Student Counseling Center

HMSU 7th Floor, 812-237-3939, www.indstate.edu/cns.

The ISU Victim Advocate

Trista Gibbons trista.gibbons@indstate.edu. HMSU 7th Floor, 812-237-3939 (office), 812-230-3803 (cell).

United Campus Ministries

321 N 7th St., Terre Haute, IN 47807 812-232-0186,

http://www2.indstate.edu/sao/campusministries.htm

http://www.unitedcampusministries.org

mailto:ucmminister2@gmail.com.

For more information on your rights and available resources:
http://www.indstate.edu/equalopportunity-titleix/titleix.

Schedule and Learning Goals

The following schedule is tentative and will be updated through out the semester.

Weeks 1 and 2, 08/20 - 08/24: Introduction

- What is Data Science?
- Learning to use Linux
- The shell - I/O redirection, pipes
- Regular expressions
- Finding things: *grep* and *find*
- Changing things: *sed* and *sort*
- Reformatting things: *sed*, *sort*, and *awk*.

Week 3 and 4, 08/27 - 08/31: Gentle review of some concepts from Statistics

- Central tendency
- Dispersion
- The normal distribution
- Laws of large numbers
- Central limit theorem
- Using R as a statistics package

Week 05, 09/03 - 09/07: More R

-

Week 05, 09/10 - 09/14: ...

-

Week 06, 09/17 - 09/21: ...

-

Week 07, 09/24 - 09/28: ...

-

Week 08, 10/01 - 10/05: ...

-

Week 09, 10/08 - 10/12: ...

-

Week 10, 10/15 - 10/19: ...

-

Week 11, 10/22 - 10/26: ...

- ...

Week 12, 10/29 - 11/02: ...

- ...

Week 13, 11/05 - 11/09: ...

- ...

Week 14, 11/12 - 11/16: ...

- ...

Week 15, 11/19 - 11/23: ...

- ...

Week 16, 11/26 - 11/30: Study week.

Week 17, 12/03 - 12/07: **Final Exam**