TSCI 5050
Introduction to Data Science
Fall 2017

CLASS DAYS and TIME: Alternate Tuesdays, 10 - Noon
CLASSROOM: LIB 2.028
COURSE DIRECTOR: Alex F. Bokov, PhD
OFFICE LOCATION and HOURS: Flexible, by appointment
EMAIL: bokov@uthscsa.edu
TELEPHONE: 210 562-4106

READ THIS DOCUMENT CAREFULLY – YOU ARE RESPONSIBLE FOR ITS CONTENTS

COURSE DESCRIPTION AND OBJECTIVES

This elective course is designed to train students to use programing languages such as R and SQL to extract, prepare, and analyze data. This course is designed to be self-contained: statistical methods and theory relevant to analyzing large datasets will be covered with the computer-related course content providing tangible applications and motivating examples. In addition, the course will include organizational skill training and best practices needed to run a successful collaboration between researchers conducting patient oriented clinical research and researchers in computational fields.

Pre-requisites – None
Semester credit hours – 1.0

Learning Objectives:

- Aim: students will gain a basic working knowledge of programming languages, using R as a case study illustrating general concepts of algorithmic reasoning.
    - Students will translate a mathematical function that they understand into the form of executable code in the R language.
    - Given an R function that contains an error, students will use debug() and browser() along with reductionist reasoning to step through the function and locate the error.
- Aim: students will gain a basic working knowledge of data extraction/transformation.
    - Given a dataset and verbal inclusion, exclusion, and sorting criteria students will express those criteria as a valid SQL statement.
    - Given a table with redundant values in certain columns, students will diagram how the data should be normalized into multiple tables in a relational database.
- Aim: students will learn how to use the general linear model and the idea that statistics is a formalization of the scientific method to make sense of new statistical concepts.
    - Given a dataset and a verbal summary of the scientific goals of the study that collected the data, students will fit a multivariable linear model using R, do an automated search for an optimal set of model terms, and for each main effect or first-order interaction in the model write a sentence explaining its practical/clinical interpretation.
    - Given the default analytic output for each term in the above model, students will give the estimated effect-size, standard error, and whether or not that effect is statistically significant.

Students will use visual methods to diagnose potential problems with the above analysis and write a paragraph that can unambiguously communicate to a statistician what was already attempted and what problems require specialist assistance.

COURSE ORGANIZATION

The main teaching modalities used in this course include:

1. Didactic
2. Hands-on tutorials

Materials:

SOFTWARE AND ONLINE ACCOUNTS (all are free):

- RStudio = https://www.rstudio.com/products/rstudio/download/
- Git = https://git-scm.com/download/win and a (free) account on https://github.com/join
- An account on http://stackoverflow.com/users/login
- SQLite Studio = http://sqlitestudio.pl/?act=download

Do not contact IMS for help with course-specific software. Please contact the course director instead, bokov@uthscsa.edu

You are encouraged bring a dataset from your own research that you are having trouble with, to use as part of the coursework. You will get a lot more out of the course that way.

Homework:

Homework is assigned on a topic *before* that topic is covered. That is deliberate, because it will help you absorb the material better after having already tried to do it on your own.

COMPUTER REQUIREMENTS

Students are required to have a laptop computer that can connect to and operate over a wireless network. All laptops will connect to The UTHSCSA network via the HSCwave broadcast wireless connection. Authentication for wireless use is based on The UTHSCSA domain username and password.

Verification of proper operation prior to the start of class is highly recommended. Assistance with networking is available thru the IMS Service Desk

- Telephone: 567-7777
- E-mail ims-servicedesk@uthscsa.edu)

Assistance is also available at the IMS Student Support Center (ALTC 106). But again, only bug them about standard UTHSCSA software, not the weird course-specific stuff listed above.

Reading Assignments – Reading assignments will be listed in the individual class sections of this syllabus.

ATTENDANCE

Attendance at scheduled classes and examinations is crucial to meeting course objectives. Therefore, regular attendance in class is expected of each student.

- Attendance is defined as being present within 15 minutes after the scheduled beginning of the class and until 15 minutes before the scheduled ending of the class.
- Excused absences may be granted by the Course Director in cases such as formal presentations at scientific meetings, illness, or personal emergency.
- Excused absences are considered on an individual basis and require electronic communication with the Course Director to request an excused absence. The e-mail request to the Course Director for consideration of an excused absence must provide details regarding the circumstances and specific dates.
- It is expected that students will provide *advanced notice* of absence for scheduled events.

- If a student has excessive unexcused absences in a given course, they will automatically receive a grade of *unsatisfactory* unless *makeup* has been approved by the Course Director.
- Makeup of absences (both excused and unexcused) is allowed at the discretion of the Course Director.
- Allowable unexcused absences will be determined by the credit hours of the course as follows:

Course Semester Credit Hours	Allowable Unexcused Absences
3.0	3
2.0	2
1.0	1

TEXTBOOKS

Strictly Optional:

- Reading materials will be provided. The following are good books to own, but not required for the class:
  - Dalgaard P. Introductory statistics with R. 2nd ed. New York: Springer; 2008.
    - ISBN: 978-0-387-79053-4
    - Description: Light and easy.
  - Burns P. S poetry. Lulu Com; 2012.
    - ISBN: 978-1-4710-4552-3
    - Description: Small but wise. Like Yoda.
  - Crawley MJ. The R book. Chichester, England?, Hoboken, N.J: Wiley; 2007.
    - ISBN: 978-0-470-51024-7
    - Description: Heavier, more detailed.
  - Pinheiro JC, Bates DM. Mixed-effects models in S and S-PLUS. New York: Springer; 2000.
    - ISBN: 978-0-387-98957-0
    - Description: Reality check for the over-confident.

GRADING POLICIES AND EXAMINATION PROCEDURES

Satisfactory/Unsatisfactory based on best 4 assignments.

Grading System

The grading will be conducted on a pass fail basis and both assignments need a Satisfactory in order to pass the course. S = Satisfactory U = Unsatisfactory

REQUESTS FOR ACCOMODATIONS FOR DISABILITIES

In accordance with policy 4.2.3, **Request for Accommodation Under the ADA and the ADA Amendments Act of 2008 (ADAAA)**, any student requesting accommodation must submit the appropriate request for accommodation under the American with Disabilities Act (ADA, form 100). To his/her appropriate Associate Dean of their School and a copy to the ADA Coordinator. Additional information may be obtained at http://uthscsa.edu/eeo/request.asp.

ACADEMIC INTEGRITY AND PROFESSIONALISM

Any student who commits an act of academic dishonesty is subject to discipline as prescribed by the UT System Rules and Regulations of the Board of Regents. Academic dishonesty includes, but is not limited to, cheating, plagiarism, collusion, the submission for credit of any work or materials that are attributable in whole or in part to another person, taking an exam for another person, signing attendance sheets for another student, and any act designed to give unfair advantage to a student or the attempt to commit such an act. Additional information may be obtained at http://catalog.uthscsa.edu/generalinformation/generalacademicpolicies/academicdishonestypolicy/

TITLE IX AT UTHSCSA

Title IX Defined:

Title of the Education Amendments of 1972 is a federal law that prohibits sex discrimination in education. It reads "no person in the United States shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity receiving Federal financial assistance."

University of Texas Health Science Center San Antonio's Commitment:

University of Texas Health Science Center San Antonio (UTHSCSA) is committed to maintaining a learning environment that is free from discriminatory conduct based on gender. As required by Title IX, UTHSCSA does not discriminate on the basis of sex in its education programs and activities, and it encourages any student, faculty, or staff member who thinks that he or she has been subjected to sex discrimination, sexual harassment (including sexual violence) or sexual misconduct to immediately report the incident to the Title IX Director.

In an emergency, victims of sexual abuse should call 911. For non-emergencies, they may contact UPD at 210-567-2800. Additional information may be obtained at http://students.uthscsa.edu/titleix/

EMAIL POLICY

All correspondence will be sent to the student using the student's LiveMail address. All correspondence from the student to the course director should be sent to the course director's e-mail as listed on the first page of this syllabus, *not* CANVAS.

USE OF RECORDING DEVICES

Encouraged. Please send course director a copy!

ELECTRONIC DEVICES

Cell phones must be put on silent mode during all class meetings and exams. Computers and other devices should be used only for participating in classroom activities.

TENTATIVE CLASS SCHEDULE

TSCI 5050
Introduction to Data Science
Fall 2017

Week	Date	Module	Title/Instructor(s)
1	August 22	**Introduction**	Introduction and GitHub
2	September 5	**R**	Anatomy of a Programming Language, Part 1
3	September 19	**R**	Anatomy of a Programming Language, Part 2
4	October 3	**SQL**	Data Storage and Manipulation
5	October 17	**Stats**	Stats Crash Course 1, (almost) Everything is Some Form of Regression
6	October 31	**Stats**	Stats Crash Course 2, Real Life Is Never Tidy
7	November 14	**Stats**	Stats Crash Course 3, Survival Analysis and Time Series
8	November 28	**Wrap-Up**	Special Topics: Data Visualization
9	December 12	**Wrap-Up**	Special Topics: Questions, Make-Up

Week: 1

Date: August 22

Room: LIB 2.024

Instructor(s): Bokov

Topic: Introduction and GitHub Tutorial

Learning Objectives and Competencies– Participants will be able to:

1. Understand structure and expectations of the course; introductions
2. Articulate what goals you might wish to achieve using biomedical informatics
3. Share any relevant research projects and experiences with the class
4. Be able to use git to clone, checkout, commit, push, and merge. And explain what each of them does.

Class Assignment:

1. Before and during first class: install software needed for course
2. Problem set: discover intent of unfamiliar samples of R code
3. In one or two sentences, describe a data problem you face in your own work that you hope to solve computationally

Readings:

http://blog.scottlowe.org/2015/01/14/non-programmer-git-intro/

Sections from free e-book S-Poetry indicated in the homework #1 handout.

Week: 2

Date: September 5

Room: LIB 2.028

Instructor(s): Bokov

Topic: Anatomy of a Programming Language, Part 1

Learning Objectives and Competencies– Participants will be able to:

1. Understand the features that almost all programming languages have in common
2. Be able to perform basic arithmetic, flow-control, and convenience/utility tasks in R
3. Know where to go to find answers to R and what other resources exist online for improving your skills as you go along.
4. Be able to turn a script into a function.
5. Be able to create an empty function and write it from the “inside out” using the browser() command

Class Assignment:

1. May opt to revise pre-class exercise #2
2. Narrow the scope of pre-class exercise #3 and attempt to write an R function that is useful to this task.

Readings:

TBA

Week: 3

Date: September 19

Room: LIB 2.028

Instructor(s): Bokov

Topic: Anatomy of a Programming Language, Part 2

Learning Objectives and Competencies– Participants will be able to:

1. Use the lapply() and sapply() functions instead of loops.
2. Use the %>% operator to keep nested function calls readable.
3. Demonstrate the use of all the subscripting operators.
4. Use the split() command.

Class Assignment:

1. May opt to further revise pre-class exercise #2
2. Use what you learned to improve the function you wrote for exercise #3 or write a new function for some different purpose.
3. Homework #2: write SQLite scripts demonstrating use of SELECT, WHERE, GROUP BY, and JOIN.

Readings:

Read TBA sections from SQL text, highlighting the parts you don't understand and bring marked-up printout to class.

Week: 4

Date: October 3

Room: LIB 2.028

Instructor(s): Bokov

Topic: Data Storage and Manipulation

Learning Objectives and Competencies– Participants will be able to:

1. Know the types problems SQL was designed to solve, and how to determine whether SQL is the right tool for solving a given problem.
2. Be able to demonstrate the use of column and row criteria via the SELECT and WHERE clauses
3. Be able to explain what a relational database is and demonstrate the use of relational logic via JOIN clauses.

Class Assignment:

Homework #3

Readings: TBA

Week: 5

Date: October 17

Room: LIB 2.028

Instructor(s): Bokov

Topic: Stats Crash Course 1, (almost) Everything is Some Form of Regression

Learning Objectives and Competencies– Participants will be able to:

1. Be able to explain what a test statistic is, what a statistical distribution is, what they have to do with hypothesis tests, and what p-values actually are.
2. Be able to use visual analogies to explain the relationship between T-tests, ANOVA, and various types of regression.
3. Know the commands for fitting a simple linear regression model in R and interpret the most important parts of the output it yields.
4. Demonstrate basic skills for getting some benefit out of a quantitative journal article without necessarily understanding all the proofs.

Class Assignment:

Apply the steps learned in class to update the analysis done in Homework #3

Readings:

Pick one from a TBA selection of peer-reviewed journal articles, critique use of statistics, and highlight the parts you don't understand

Week: 6

Date: October 31

Room: LIB 2.028

Instructor(s): Bokov

Topic: Stats Crash Course 2, Real Life Is Never Tidy

Learning Objectives and Competencies– Participants will be able to:

1. Integrate the material learned from the SQL and R units to understand what analysis-ready data should look like.
2. Be able to use residual plots and other diagnostic methods to validate your analysis, identify problems, and know what to do about them.
3. Be able to use contrasts to convert the default hypothesis tests of a regression model into ones that are relevant to your actual hypotheses.
4. Understand how to effectively collaborate with statisticians.

Class Assignment:

1. Perform diagnostic procedures learned in class on the analysis done in Homework #3
2. Attempt remedial measures where applicable

Readings:

TBA

Week: 7
Date: November 14
Room: LIB 2.028
Instructor(s): Bokov
Topic: Stats Crash Course 3, Survival Analysis and Time Series
Learning Objectives and Competencies– Participants will be able to: 1. Be able to show examples other than mortality of where survival analysis is an appropriate approach 2. Know the commands for fitting a simple survival model in R and interpret the most important parts of the output it yields 3. Know how to perform diagnostics on survival models and what to do about them.
Class Assignment: If a time-to-event approach makes sense for the dataset you have been working on in previous homeworks, try this approach. Otherwise, use one of the example datasets that come with R. Prepare final questions to ask next class.
Readings: TBA

Week: 8
Date: November 28
Room: LIB 2.028
Instructor(s): Bokov
Topic: Wrap-Up: Data Visualization
Learning Objectives and Competencies– Participants will be able to: 1. Be able to construct a ggplot object and quickly change the type of plot it renders by adding additional layers. 2. Be able to create a dynamically updatable report using R markdown.
Class Assignment: ggplot exercise, optionally, Shiny tutorial
Readings: None

Week: 9
Date: December 12
Room: LIB 2.028
Instructor(s): Bokov
Topic: Wrap-Up
Learning Objectives and Competencies– Participants will be able to: 1. Question and answer session.
Class Assignment: Have a great summer and keep in touch!
Readings: None