TSCI 5050: Introduction to Data Science **1 Hrs** **Fall 2016**

Instructional Team:
Alex F. Bokov Ph.D. (Director), Donald C. McCurnin M.D., Sandun Jayarathna Pahula Hewage Ph.D., Yufan Zhou Ph.D.

LEARNING OBJECTIVES and COMPETENCIES

- 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.

All Classes Held Alternating Tuesdays, 10:00am – 12:00am in AL&TC 2.211

Notes:

- Please bring a laptop to every class.
- Please set your phone on vibrate.
- For most class sessions, the homework will be assigned during the *previous* class and reviewed *in* the next class. I.e. do not panic if you encounter certain concepts for the first time in your homework. Write down your questions and they will be addressed in class.
- Alex F. Bokov's email address is bokov@uthscsa.edu

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).

SOFTWARE REQUIRED:

- RStudio = https://www.rstudio.com/products/rstudio/download/
- Python = https://www.python.org/downloads/
- 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 installing course-specific software. Please contact the course director instead.

GRADING POLICY:

Satisfactory/Unsatisfactory based on best 4 assignments.

ATTENDANCE POLICY:

- 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

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/

Introduction to Data Science
Week: 1
Date: 8/23/2016
Topic: Introduction
Instructors: Alex F. Bokov
Learning Objectives: 1. Understand structure and expectations of the course; introductions 2. Summarize the three domains of biomedical informatics 3. Articulate what goals you might wish to achieve using biomedical informatics 4. Share any relevant research projects and experiences with the class
Class Assignment: 1. Install R-studio and SQLite viewer 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 and Bibliography: Read printouts of sections in “S Poetry”, highlighting the parts you don't understand (marked-up printouts will be collected)

Introduction to Data Science
Week: 2
Date: 9/6/2016
Topic: R and Python, Anatomy of a Programming Language
Instructors: Alex F. Bokov, Sandun
Learning Objectives: 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 questions, and what other resources exist online for improving your R skills as you go along. 4. Understand the workflow of software engineering and how to effectively collaborate with programmers.
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. 3. Problem set: write SQLite scripts demonstrating use of SELECT, WHERE, GROUP BY, and JOIN.
Readings and Bibliography: Read TBD sections from a text about SQL, highlighting the parts you don't understand (marked-up printouts will be collected)

Introduction to Data Science
Week: 3
Date: 9/20/2016
Topic: SQL, the Language of Data
Instructors: Alex F. Bokov
Learning Objectives: 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: 1. May opt to revise pre-class exercise #3 2. Complete online HIPAA training, fill out system access agreement for data warehouse. 3. Once account created by CIRD, develop some i2b2 queries related to your research topic.
Readings and Bibliography: Watch the introductory video on i2b2.org

Introduction to Data Science
Week: 5
Date: 10/18/2016
Topic: Stats Crash Course 1, (almost) Everything is Some Form of Regression
Instructors: Alex F. Bokov
Learning Objectives: 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: 1. Apply the steps learned in class to update the analysis done in Week 4 homework.
Readings and Bibliography: Pick one from a TBD selection of peer-reviewed journal articles, critique use of statistics, and highlight the parts you don't understand (marked-up printouts will be collected)

Introduction to Data Science
Week: 6
Date: 11/1/2016
Topic: Stats Crash Course 2, Real Life Is Never Tidy
Instructors: Alex F. Bokov
Learning Objectives: 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 Week 5 homework 2. Attempt remedial measures where applicable 3. One-on-one meetings simulating initiation of a collaboration with a statistician
Readings and Bibliography: TBD

Introduction to Data Science
Week: 7
Date: 11/15/2016
Topic: Stats Crash Course 3, Survival Analysis, Time Series, and Clinical Practice in the Neonatal ICU
Instructors: Donald C. McCurnin and Alex F. Bokov
Learning Objectives: 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: 1. 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.
Readings and Bibliography: TBD

Introduction to Data Science
Week: 8
Date: 11/29/2016
Topic: Data Mining – Clustering, Decision Trees, Visualization
Instructors: Alex F. Bokov
Learning Objectives: 1. Understand the type of clustering algorithms that exist along with their benefits and drawbacks. 2. Demonstrate how to create and validate a decision tree, and be able to discuss the broader problem of cross-validation. 3. Demonstrate the use of R's rich visualization libraries and how they can be integrated into an online dashboard app. 4. Understand the continued role of “traditional” hypothesis testing in the brave new world of “big data”
Class Assignment: 1. Determine whether your patient cohort from previous homeworks breaks down into any plausible clusters. Are there any variables that can explain this clustering? 2. Use some of the visualization methods covered in class on your dataset. Incorporate this and results of previous homework to prepare a two-page research plan and methods targeted at a grant reviewer with no computational or statistical background. 3. Know the mapping tools and transcriptome analysis tools and their parameters. (TBD)
Readings and Bibliography: 1. Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science, 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369. 2. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods, 2012 Oct;9(10):999-1003. doi: 10.1038/nmeth.2148. Epub 2012 Sep 2.

Introduction to Data Science
Week: 9
Date: 12/13/2016
Topic: Next-Gen Sequencing and Hi-C Sequencing
Instructors: Sandun Jayarathna Pahula Hewage Ph.D., Yufan Zhou Ph.D.
Learning Objectives: 1. General introduction to Hi-C. 2. Demonstrate a data analysis pipe-line in Python used in Hi-C. 3. Introduction to PCA (Principal Component Analysis) and chromosomal compartments. 4. Integration of epigenetic data with Hi-C 5. Understand the production of next generation sequencing data. 6. Know how to map reads to a large reference genome such as the human genome. 7. Be able to do transcriptome analysis.
Class Assignment:
Readings and Bibliography: .