STA 327: Machine Learning, Data Mining, and Information Retrieval
Duke University, Fall 2020

Instructor: Rebecca Steorts, Assistant Professor, Dept of Statistical Science, beka@stat.duke.edu
Course: Tu/Th, 1:25 - 2:40 PM (116 Old Chemistry)
Lab: Monday: 1:25 - 2:40 PM (Sociology Psychology 130)
Steorts Office Hours: Tuesday/Thursday: 2:40 – 3:40 PM, Office: Old Chemistry 216.

Teaching Assistant: Ben Feder, benjamin.feder@duke.edu
Lab: Monday: 1:25 - 2:40 PM (Sociology Psychology 130)
Office Hours: Wednesday 6:30-8:30 PM (Old Chemistry 203)

Course description: The rapid growth of digitalized data and the computer power available to analyze it has created immense opportunities for both machine learning and data mining. This course introduces exploratory data analysis, functional programming, machine learning, and data mining methods. Topics covered include information retrieval, clustering, classification, modern regression, cross validation, boosting, and bagging. The course emphasizes selection of appropriate methods and justification of choice, use of programming for implementation of the method, and evaluation and effective communication of results in data analysis reports.

The course has the following learning objectives:

1. Writing functional programming
2. Learning how to write reproducible code
3. Learning how to work with github
4. Performing exploratory data analysis
5. Working in collaborative groups and teams
6. Learning the fundamentals of data mining/information retrieval
7. Learning when to apply different data mining/information retrieval methods
8. Evaluation/Presenting findings of data mining/information retrieval methods

Finally, students will put all the above *learning objectives* together in a hackathon, where they will work in small groups to attack a real data set, in real time. The goal will be to analyze a real data set using data mining methods they have seen or new methods they have not seen, and write a report on their findings/conclusions. Students will have time after the datathon to revise their reports, and then give a presentation, which will serve as the final project.

Course webpage: `https://github.com/resteorts/data-mine`

Labs: The lab for this course is optional, but the lab section is dedicated for extra sessions that will be held by the TA to review lab and applied exercises that are a part of the in-course material or ones that are in the ISLR book. You are encouraged to attend these if you're having trouble keeping up in class for any reason. Make up classes may be held during this extra lab period as well. These will be announced in advance.

Prior Knowledge: Students are expected to have a solid background in regression analysis (STA 210) and elementary probability (STA 230). These will be building blocks for course topics, and little review will be provided during class time. If you are unsure what prior knowledge is expected, please refer to the following past syllabi:

1. STA 210: `https://www2.stat.duke.edu/courses/Spring19/sta210.001/`
2. STA 230: `https://www2.stat.duke.edu/cour`

Expectations: Students are expected to be very familiar with `R` and will be expected to have learned `R` markdown by the end of the course. All homeworks, reports, and take home exams (if applicable) should be submitted in Markdown `.Rmd` and `.pdf` format. Please name your reports using your net id. Your reports are expected to be reproducible and compile for full credit. Students are expected to keep up with the reading in the course and have read before they come to class. Finally, if students find typos on the slides, please write them down with the slide and typo and give them to Professor Steorts for a timely correction to the course webpage.

Re-grades: If you believe that you lost points on a homework, please write an email to the instructor and the TA for the course explaining why you think you lost points in the assignment, and your re-grade request will be considered. Re-grades must be considered in writing.

Recommendations: It is recommended that students work through all lab exercises in the required text book. Additional homework will be given the supplement this. Again, it is recommended that students read in advance of the material covered in class. It is also recommended that students attend office hours if they are having trouble early in the semester. We are here to help you all and answer questions!

Required Texts:
An Introduction to Statistical Learning with Applications in `R`, Gareth James, Daniela Whitten Trevor Hastie, and Robert Tibshirani, (2013), Springer.

Supplemental Texts:
The Art of `R` *Programming: The Tour of Statistical Software Design*, Norman Matloff (2011). `http://diytranscriptomics.com/Reading/files/The%20Art%20of%20R%20Programming.pdf`

Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Ed., Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). `http://statweb.stanford.edu/~tibs/ElemStatLearn/`

The R Cookbook, `http://www.cookbook-r.com/`.
Mining of Massive Datasets. Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman (2010). `http://infolab.stanford.edu/~ullman/mmds/book.pdf`.

Table 1: Grading Policy:

Homework	30%
Exam I (Tuesday, September 25)	20%
Duke Datathon (Saturday, October 27)	30%
Exam II (Thursday, November 29)	20%

Table 2: Grading Scale:

90 – 100	A
80 – 89	B
70 – 79	C
60 – 69	D
0 – 59	E

Exams: Exams will be closed book and closed notes.

Course Policies: Homework assignments will be announced on Sakai (along with the due date). Late homeworks will not be accepted. Your lowest homework grade will be dropped to take in to consideration things that arise during the semester.

Homework expectations: All homework's involving analysis and code must be submitted to Sakai using Markdown and RStudio. Specifically, your homework must be reproducible. Your homework must be included as one file, therefore, please zip your files and submit all the files using a .zip extension. If you are unsure of how to do this, please see your TA. Submissions via email to the TA's or instructor will not be accepted for credit. Please submit early and often. Again, late submission will not be accepted.

Homework derivations: Please note that derivations for homework can be submitted in any format of your choosing as long as your convert this to a pdf file. Please also note that your work must be legible to myself and the TA's.

Duke Datathon: Duke Datathon will be a required event for our class this year. You will work in teams of groups of three students during this event to solve a challenging problem. See `http://dukeml.org/datathon/` for more information. After Datathon, you will create a 6 minute presentation with your group to present to the class on your Datathon experience. You will also turn in a three page report on your findings and include all your code in the Appendix of the report. Details of this will be discussed in class, and this component will count towards your final grade.

Discussion board: There is a Google course discussion page. Please direct questions about homeworks and other matters to that page. Otherwise, you can email the instructors (TAs and professor). Note that we are more likely to respond to the Google questions than to the email, and your classmates may respond too, so that is a good place to start. You can ask for permission to add to the group at https://groups.google.com/forum/#!forum/data-mine18.

Missing class/exams/work: You are responsible for everything from lecture. Do not depend on the course web page for announcements regarding due dates for homework, changes in schedules, etc. Such announcements will be made in class. Homework assignments will be uploaded to the course webpage along with course readings (please check here frequently for updates).

Students who miss graded work due to a scheduled varsity trip, religious holiday or short-term illness should fill out an online NOVAP, religious observance notification or short-term illness notification form, respectively. If you are faced with a personal or family emergency or a long-range or chronic health condition that interferes with your ability to attend or complete classes, you should contact your academic dean's office. See more information on policies surrounding these conditions here https://trinity.duke.edu/undergraduate/academic-policies/personal-emergencies, and your academic dean can provide more information as well.

Makeup exams must be approved before the time of the exam and will be given only in case of medical or family emergencies (which must be appropriately documented – see above). All work turned in for a grade must be entirely your own. This particularly relates to homework. You are encouraged to talk to each other regarding homework problems or to the instructor/TA, however the write up, solution, and code *must* be entirely your own solution and work.

Academic Honesty: Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. Cheating on exams and quizzes, plagiarism on homework assignments, projects, and code, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved as well as being reported to the University Judicial Board. Additionally, there may be penalties to your final class grade. Please review Duke's Standards of Conduct. For more information on the Duke honor code (known as Duke Community Standard), please go to http://integrity.duke.edu/faq/faq1.html.

Cell phones and laptops: Cell phones should be turned off (or set on silent). Also, please try and be courteous of other students if you bring a laptop or food to class.

Students with Disabilities: Students who require special accommodations in class or during exams should follow the procedures outlined by the Disability Management Program http://access.duke.edu/students. Students with disabilities who believe they may need accommodations in this class are encouraged to contact the Student Disability Access Office at (919) 668-1267 as soon as possible to better ensure that such accommodations can be made. If you have a special accommodation outlined by the Disability Management Program, I encourage you to please meet with me during the first week of class. Please email me to set up an appointment or speak with me in person.

Privacy Policies: Student records are confidential. For example, it is again Duke policy for any instructor/TA to email your grades to you.
Proposed Topics (Subject to Change):

1. What is machine learning and data mining?
2. Exploratory Data Analysis
3. Functional Programming
4. Reproducible Code and Working with Repositories
5. An Introduction to Information Retrieval
6. Locality Sensitive Hashing
7. Locality Sensitive Hashing and Entity Resolution
8. Clustering