# STAT 380: Data Science Through Statistical Reasoning and Computation

Spring 2019

# Contents

purse Information
Teaching Team
Prerequisites
Class Time & Location
Websites & Course Links
Laptops & Textbooks
Communication
Grading
Course Goals and objectives
blicies & Resources
Counseling and Psychological Services (CAPS)
ECoS Code of Mutual Respect
Academic Integrity Statement
Disability Policy
ellabus Changes 6

# **Course Information**

# Office Hours

Day	When	Where	Who
Tues	2:00-3:00 PM	423 Thomas Bldg	Ling Zhang
Thurs	10:30-11:30 AM	323 Thomas Bldg	Beckman
Thurs	2:00-3:00 PM	423 Thomas Bldg	Ling Zhang
Friday	2:30-3:30 PM	323 Thomas Bldg	Beckman
Friday	5:00-6:00 PM	330 Thomas Bldg	Alex Chen

# **Teaching Team**

#### Instructor

Dr. Matthew Beckman office: 323F Thomas Building phone: (814) 863-1022 PSU email ID: mdb268@psu.edu

# **Teaching Assistants**

Alex Chen (auc39@psu.edu)

Ling Zhang (luz136@psu.edu)

# Prerequisites

STAT 184 & STAT 200  $\,$ 

# Class Time & Location

Time/Day: MWF 1:25 - 2:15pm Location: 144 Stuckeman Family Bldg

# Websites & Course Links

Canvas: https://psu.instructure.com Course Resouce Page: https://mdbeckman.github.io/PSU-STAT-380/

# Laptops & Textbooks

# Laptops

Bring a laptop to class each day if you have one. I encourage students to work collaboratively, so I'd like to have at least one laptop for every 2 or 3 students. There's no need to buy a laptop if you don't already have one.

# Textbooks

(required): Modern Data Science with R by Benjamin Baumer, Daniel Kaplan, Nicholas Horton

(required): Advanced R by Hadley Wickham (free ebook: http://adv-r.had.co.nz/)

Other materials (e.g. book chapters, articles) may be distributed as needed during the term

# Communication

# Piazza

We will be using Piazza for out of class Q&A, to help you benefit from each other's questions and the collective knowledge of your classmates, professor, TA. Questions should be posted to the entire class. I encourage you to ask questions if you are struggling to understand a concept, and to answer your classmates' questions when you can.

**Do NOT** use Piazza for issues related to your grade or other private matters (not even an "instructor post"); email those questions or comments to the professor/TA directly or discuss them in person.

# Email

Most issues about classroom activities (content, logistics, etc) should be posted to Piazza, but you should email (or have a conversation in person) for personal or private matters (e.g. grades, illness).

#### Grading

Learning outcomes will be assessed based on performance in each of the following categories accompanied by their impact on the overall grade:

#### Tentative Percentages 35% Exams

25% Project(s)20% Programming notebooks 15% Exercises (e.g., homework sets, in-class tasks)3% Participation

#### **Missing Class**

Students who miss class are responsible for the content covered. At times, unannounced tasks might be assigned and completed during class without warning. These cannot be made-up or submitted late.

#### Late Work

Most assignments are turned in electronically on Canvas, so the official time stamp on Canvas will be used to determine late work. Basically, there is no "on time" with Canvas, so even 8:00:00.001am is technically *after* 8:00am. In Canvas, as the cliche goes: "If you're not early, you're late."

Credit for late work will generally be awarded as follows:

- 80% credit for the first 24 hours after the due time/date;
- 50% credit between 24 & 48 hours after the due time/date;
- Zero credit awarded after 48 hours beyond the due date/time

#### Final grades

Grade	Score
A	> 93%
A-	90%
B+	87%
В	83%
B-	80%
C+	77%
С	70%
D	60%
F	< 60%

#### Exams

There will be a midterm exam and a final exam. The format of each is TBD

## **Project**(s)

One or more projects will be assigned during the course of the term. The format and expectations may vary somewhat widely among projects.

#### **Programming Notebooks**

Students will regularly be asked to program through sections/chapters of the book in order to gain practice with the content and learn by reproducing expert code provided by the authors. Programming Notebooks are

expected to conform to the R Style Guide for class, and generally should reproduce all code and accompanying output in the assigned section.

# Participation

Participation is graded in two parts based largely on Piazza contributions: (A) Community; (B) Self-sufficiency.

A. Community. Students can earn one point for each question posed to the class, and 2 points for each question answered up to the maximum point total in Canvas. For example, if the assignment is worth 30 points, students could earn full credit by asking 10 questions (1 pt ea.) and answering 10 questions (2 pts ea.) during the term. Grading will utilize Piazza meta-data that can be accessed only by an instructor. Notes: (1) Piazza only credits you with an "Answer" if you edit the "Students' Answer" box... "Follow up discussions" should be used for threaded discussion; (2) unproductive posts may be removed by an instructor or TA and will therefore not earn points.

B. Self-sufficiency. Students are also incentivized to be reasonably self-sufficient. These are important habits for a data scientist to develop, so this score is adjusted for questions (e.g. Piazza posts, emails) that could have been easily answered with an Internet search (e.g. Google) or are explicitly printed in the syllabus.

# Exercises

Assigned exercises will include some combination of homework sets from the textbook, in-class exercises (not necessarily announced in advance), and other tasks. Students are generally *encouraged* to work together on such assignments, but each student must hand in their own work (unless unstructed otherwise).

## Course Goals and objectives

Some goals and objectives may be reduced or expanded as time permits, but a tentative list follows:

- General Tools
  - R programming language
  - RStudio development environment
  - Literate programming and R Markdown
  - Git/GitHub source control
- Practice good programming style
- Demonstrate proficiency with the following
  - Intermediate data visualization using ggplot2
  - Intermediate data wrangling using tidyverse
  - Tidy data and iteration using tidyverse
    - \* reshaping data
    - \* joins
    - \* vectorized operations
    - \* iteration (e.g., dplyr::do and mosaic::do)
    - \* intermediate data intake
  - Basic statistical foundations in R
  - Supervised learning & predictive analytics
    - \* Supervised learning
    - \* Classifiers (e.g., decision trees, random forests, naive Bayes, neural networks)
    - \* Ensemble methods
    - \* Model Evaluation
    - \* Regularization
  - Unsupervised learning
    - \* Clustering
    - \* Dimension reduction

- Simulation
- Interactive graphics (e.g., Leaflet, Plot.ly, ggvis, Shiny)
- Database querying using SQL
- Working with spatial data (e.g., static & dynamic maps, projections, geocoding, routes, distances)
- Text as data
  - \* Regular expressions
  - \* Ingesting text
  - \* Analyzing textual data
- Network science
- Notions & tools for "big data"

# Policies & Resources

# Counseling and Psychological Services (CAPS)

Many students at Penn State face personal challenges or have psychological needs that may interfere with interfere with their academic progress, social development, or emotional wellbeing. The university offers a variety of confidential services to help you through difficult times, including individual and group counseling, crisis intervention, consultations, online chats, and mental health screenings. These services are provided by staff who welcome all students and embrace a philosophy respectful of clients' cultural and religious backgrounds, and sensitive to differences in race, ability, gender identity and sexual orientation.

Counseling and Psychological Services at University Park (CAPS):

- Phone: 814-863-0395
- Web: http://studentaffairs.psu.edu/counseling/

## Penn State Crisis Line (24 hours/7 days/week): 877-229-6400

## Crisis Text Line (24 hours/7 days/week): Text LIONS to 741741

## ECoS Code of Mutual Respect

The Eberly College of Science Code of Mutual Respect and Cooperation embodies the values that we hope our faculty, staff, and students possess and will endorse to make the Eberly College of Science a place where every individual feels respected and valued, as well as challenged and rewarded.

#### Academic Integrity Statement

Academic dishonesty is not limited to simply cheating on an exam or assignment. The following is quoted directly from the "PSU Faculty Senate Policies for Students" regarding academic integrity and academic dishonesty:

Academic integrity is the pursuit of scholarly activity free from fraud and deception and is an educational objective of this institution. Academic dishonesty includes, but is not limited to, cheating, plagiarizing, fabricating of information or citations, facilitating acts of academic dishonesty by others, having unauthorized possession of examinations, submitting work of another person or work previously used without informing the instructor, or tampering with the academic work of other students.

All University and Eberly College of Science policies regarding academic integrity/academic dishonesty apply to this course and the students enrolled in this course. Refer to the following URL for further details on the academic integrity policies of the Eberly College of Science: http://www.science.psu.edu/academic/Integrity/ index.html. Each student in this course is expected to work entirely on her/his own while taking any exam,

to complete assignments on her/his own effort without the assistance of others unless directed otherwise by the instructor, and to abide by University and Eberly College of Science policies about academic integrity and academic dishonesty. Academic dishonesty can result in assignment of "F" by the course instructors or "XF" by Judicial Affairs as the final grade for the student.

# **Disability Policy**

Penn State welcomes students with disabilities into the University's educational programs. If you have a disability-related need for reasonable academic adjustments in this course, contact Student Disability Resources (SDR; formerly ODS) at 814-863-1807, 116 Boucke, http://equity.psu.edu/student-disability-resources. In order to receive consideration for course accommodations, you must contact ODS and provide documentation (see the guidelines at http://equity.psu.edu/student-disability-resources/guidelines).

# Syllabus Changes

This syllabus is subject to change as circumstances warrant; all changes will be distributed in writing (usually electronically with Canvas, Class Webpage) or announced in class.

This document was last modified on January 11, 2019