One special aspect of this course is that it will be team-taught with several instructors. The schedule and content sequencing is described herein.

All of the contributing faculty are:

Amanda Hering	Statistics	Baylor	mandy_hering@baylor.edu
Tzahi Cath	Environmental Engineering	CSM	tcath@mines.edu
Doug Nychka	Statistics	CSM	nychka@mines.edu
Michael Poor	Computer Science	Baylor	michael_poor@baylor.edu
Greg Hamerly	Computer Science	Baylor	greg_hamerly@baylor.edu

Graduate students and TAs are:

Aurora Waclawski	Environmental Engineering	CSM	awaclawski@mymail.mines.edu
TA: Maggie Bailey	Statistics	CSM	mdbailey@mymail.mines.edu
TA: Luke Durell	Statistics	Baylor	Luke_Durell1@baylor.edu

Office Hours are as follows: We encourage you to contact the professor whose specific lecture or in-class assignment you have a question about. All times are in Central Standard Time.

Amanda Hering	Thursday 9-10 am	`https://baylor.zoom.us/j/7720784879?pwd=WXlSZ2dab04yU3VRVlhCY29tZnN`
Doug Nychka	Wednesday 11-12 pm	`https://mines.zoom.us/j/5797557040`
Michael Poor	Tuesday 1-3 pm	Contact him through Microsoft Teams
Greg Hamerly	Wednesday 3-4 pm	`https://baylor.zoom.us/j/4808730894?pwd=OFUwUmlja2huSTJOVzdiOGtzUzN`
Luke Durell	Friday 11-12 pm	`https://baylor.zoom.us/j/83736537266?pwd=WVA2dndZNmp3Y05OV3dRQnhzcG`

Course Description: Introduction to principles of data science, including problem workflow, variable types, visualization, modeling, programming, data management and cleaning, reproducibility, and big data.

Prereq: None!!

Course Schedule: T/Th 11:00-12:15 pm

Course Materials: Materials for the course can be found on Canvas.

Hardware/Software: You are expected to have a laptop that can run RStudio. You will also need a strong WiFi connection so that you can fully participate in class everyday.

Course Introduction: "Data is the sword of the 21st century, those who wield it well, the Samurai." —Jonathan Rosenberg, adviser to Larry Page and former SVP of products at Google. The quote appeared in "The Official Google Blog" on February 16, 2009.

[**What is Data Science?**] Data science is a field that combines mathematics, computer science, and statistics with the goal of answering questions and discovering new information from data. Some examples are: How should I buy a used car on cars.com? Where will the next flu outbreak occur? How is the climate changing in Texas? What is the electrical power consumption of a supercomputer? These kind of practical questions require skills in wrangling data sets into forms for analysis, using graphics to explore relationships among variables, writing programs to do data-driven analysis, and communicating the results in nontechnical language.

[**What is this course about?**] This course will expose students to data analysis and discovery using data science. In the process, students will learn how to write programs in the R language and generate figures and reports. R is a community-based data analysis environment that is free and has become one of the standard programming languages in data science and statistics. The course will introduce students to some modern data analysis tools, including regression and smoothing, multivariate analysis, clustering, databases, etc. Some of the statistical and mathematical background for the analysis techniques will be given, but the emphasis will be on solving real data problems and learning how to work with these methods. We will take the approach that many sophisticated and advanced methods can be appreciated and used within the context of particular data sets if students have a clear idea of the analysis goals and an understanding of how the data is collected or generated. This kind of understanding is a practical complement to the more mathematical development that would occur in more advanced statistics or computer science courses. *This course is designed to inspire students at an early stage in their academic careers to pursue additional coursework in data science related fields.*

One unique aspect of this course is that some of the data will come from water-related applications. Why water? According to the United Nations, providing sustainable sources of clean water when and where it is needed is one of the top challenges facing the next generation.[1] On the UN's website, it states, "Water is at the core of sustainable development and is critical for socio-economic development, energy and food production, healthy ecosystems and for human survival itself. Water is also at the heart of adaptation to climate change, serving as the crucial link between society and the environment."

[**What would come next?**] This course is closely tied to a summer research program. If you are taking this course, then we encourage you to apply to a 5-week, paid ($15/hour up to 40 hours/week) summer undergraduate research Data Science Fellows Program. The program will run from June 1, 2021 to July 2, 2021. Working in interdisciplinary teams with a faculty advisor, fellows will tackle real data science problems in the water/wastewater treatment field. See `https://www.baylor.edu/mowater/` for more information and to apply.

Course Resources: There is no one required textbook for this course. Faculty contributors will likely choose material from the following list of course textbooks and/or resources.

- (Free) Wickham, H. and Grolemund, G. (2017) R for Data Science, O'Reilly Media. `https://r4ds.had.co.nz`
- (Free) Murrell, P. (2009) Introduction to Data Technologies, CRC Press.
- (Free) Murrell, P. (2005) R Graphics, 2nd ed., CRC Press.
- (Free) James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning: with Applications in R, Springer.
- Xie, Y., Allaire, J. J., and Grolemund, G. (2018) R Markdown: The Definitive Guide, Chapman & Hall/CRC The R Series.
- Nolan, D. and Lang, D. T. (2015) Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving, CRC Press.
- Nolan, D. and Lang, D. T. (2014) XML and Web Technologies for Data Sciences with R, Springer.
- Gandrud, C. (2015) Reproducible Research with R and RStudio, 2nd ed., CRC Press.
- Xie, Y. (2015) Dynamic Documents with R and knitr, 2nd ed., CRC Press.
- Saltz, J. S. and Stanton, J. M. (2018) An Introduction to Data Science, Sage.
- Baumer, B. S., Kaplan, D. T., and Horton, N. J. (2017) Modern Data Science with R, CRC Press.

[1] `https://www.un.org/en/sections/issues-depth/water/index.html`

Student Learning Outcomes: At the conclusion of this class, students should be able to:

1. Become skillful at writing programs in R and generating reports in RMarkdown.
2. Learn how to explore and check data sets using graphics and other statistics.
3. Develop skills in manipulating and transforming a data set into more useful formats.
4. Develop skills in visualizing and summarizing univariate and multivariate data.
5. Learn how to match a data-driven method to a particular type of data set or question.
6. Develop skills for communicating the results of a data analysis.

Course Work: The grade for the course will be based on the following components:

- Homework Assignments (30%): Due approximately once per week. The in-class exercises covered on Tuesday and Thursday will be due at 5 pm on Friday of that same week. A subset of questions will be graded.
- Exams (30%): Three exams will be given outside of class through Canvas and will be available to take over a 3-day window. Each one is designed to take about 45 minutes. Each exam will cover the content from approximately one-third of the class, and each is worth 10% of the overall grade.
- Final Project (25%): Teams will be formed and will get to choose from among a few datasets and questions. Each team will formulate their response and will submit both a written report and an oral presentation.
- Participation/Teamwork (15%): Presence and engagement in class.

Course Structure: Most of the lectures for this class will be flipped. There will be a 20-30 minute video that students will be required to watch *prior* to class. A coding assignment based on the video will be presented in-class, and students will be split into small groups to work on this during the remainder of the class with assistance from the instructors and TA. Students will be called upon to share their computer screen with their group or the entire class. Attendance will be taken and will count towards the participation grade. The typical workflow will be as follows:

1. Watch video.
2. Attend class.
3. Contribute to daily class discussion question.
4. Participate in in-class coding assignment. Ask questions if you do not understand, if you encounter unusual errors, or if you have a more efficient approach to answer the question.

Homework Grading and Lateness Policies: For homework assignments:

- We will allow one dropped homework assignment for the entire class if we have close to total class participation in the online course evaluation. Participation of the entire class minus one or two students is expected.
- We allow a student up to two late homework submissions, up until 8 am the morning of the following day. If you have no late homework assignment submissions by the end of the semester, then you get an extra two percentage points on each of the exams.
- No grade changes or adjustments will be made after two weeks of the graded score being returned to the student.

Grading Scale: The following letter grades are guaranteed:

Letter	Numeric Range	Letter	Numeric Range
A	90-100	C	70-72
A-	87-89	C-	67-69
B+	83-86	D+	63-66
B	80-82	D	60-62
B-	77-79	D-	57-59
C+	73-76	F	<57

Baylor University

Day	Date	Module	Primary Instructor
1	Jan 19	Intro to R	Hering
2	Jan 21	Intro to R	Hering
3	Jan 26	Intro to R	Hering
4	Jan 28	R Markdown	Nychka
5	Feb 2	R Markdown	Nychka
6	Feb 4	EDA	Hering
7	Feb 9	EDA	Poor
8	Feb 11	EDA	Poor
9	Feb 16	EDA	Hamerly
10	Feb 18	Wrangling	Nychka
11	Feb 23	Wrangling	Nychka
12	Feb 25	Guest Lecture	Cath
13	Mar 2	Programming	Hering
14	Mar 4	Programming	Poor
15	Mar 9	Graphics	Poor
16	Mar 11	Graphics	Poor
17	Mar 16	Shiny	Poor
18	Mar 18	Shiny	Poor
19	Mar 23	Regression	Hamerly
20	Mar 25	Regression	Hamerly
21	Mar 30	Variable Selection	Hamerly
22	Apr 1	Feature Generation	Hamerly
23	Apr 6	Clustering	Hamerly
24	Apr 8	Classification	Hamerly
25	Apr 13	Model Validation	Hering
26	Apr 15	GitHub	Nychka
27	Apr 20	Teamwork	Hering
28	Apr 22	Oral Presentations	Nychka
29	Apr 27	Written Reports	Nychka

Final Exam: There will be no formal final exam but rather a final project. The determination of whether presentations will be synchronous or asynchronous is TBD.