Intro to Data Science

Shawn Santo

What is Data Science?

Data science is an emerging discipline that builds on tools from mathematics, statistics, and computer science to extract knowledge from data.

To get a better understanding of data science, you will

- learn to explore, visualize, and analyze data in a reproducible and shareable manner;
- gain experience in data wrangling and munging, exploratory data analysis, data visualization, statistical inference, and predictive modeling;
- work on problems and case studies inspired by and based on real-world questions and data;
- learn to effectively communicate results through written assignments and a final project.

Some of what you will learn

- Fundamentals of `R`
- Data visualization and wrangling with `ggplot2` and `dplyr` from the `tidyverse`
- Web scraping
- Web based applications with `RShiny`
- Spatial data visualization
- Data types and functions
- Version control with `GitHub`
- Reproducible reports with `R Markdown`
- Regression and classification
- Statistical inference

Full course schedule

Teaching team

- Shawn Santo
- shawn.santo@duke.edu
- Office hours (Zoom links can be found in Sakai.)
    - Wed 12:30 pm - 1:30 pm ET
    - Thu 11:00 am - 12:00 pm ET

Teaching team

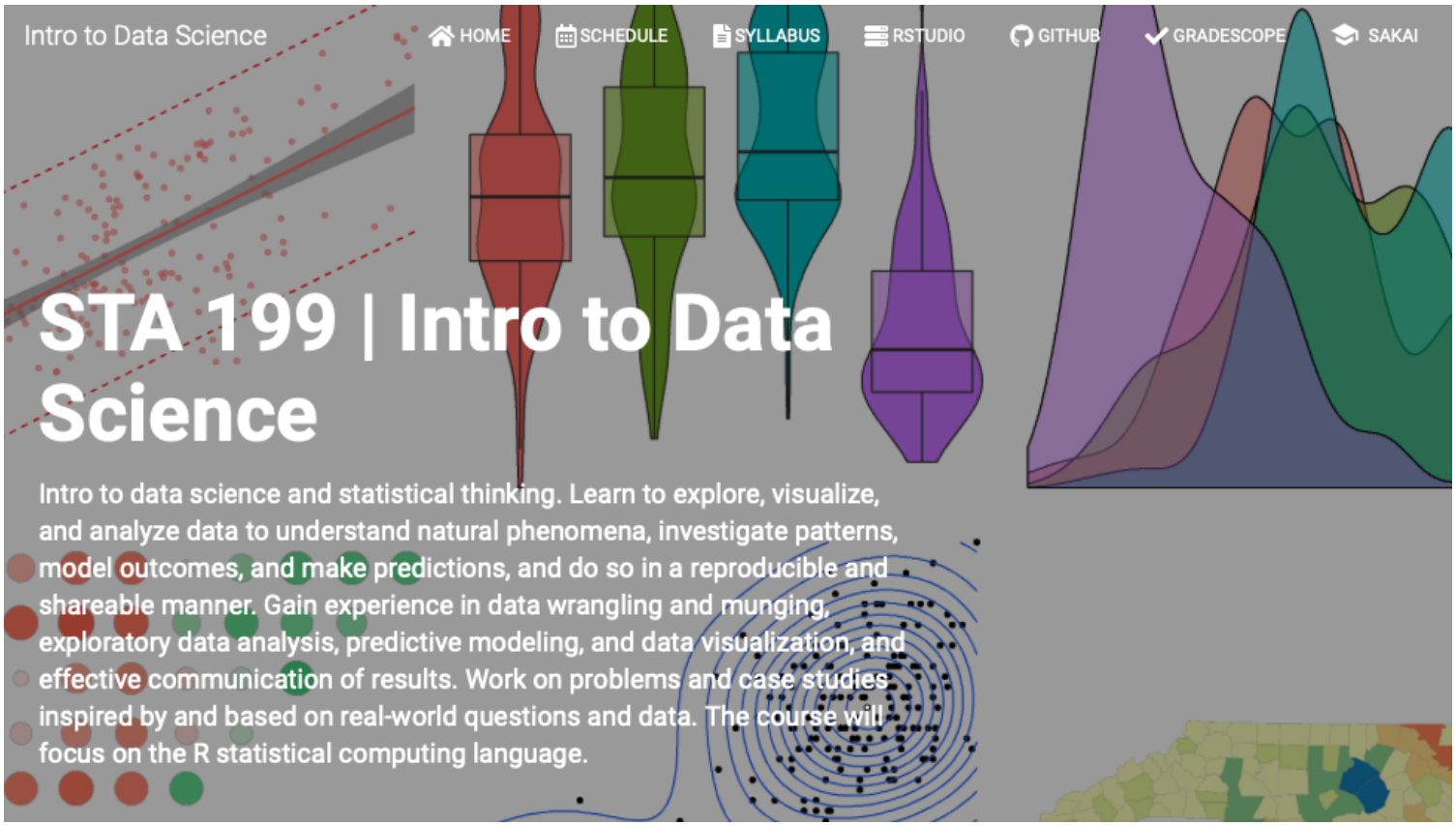
Teaching assistant	Office hours
Frances Hung	No office hours
Rob Kravec	Mon 4:00pm - 5:00pm, Wed 4:00pm - 5:00pm
Matty Pahren	Mon 2:00pm - 3:00pm, Thu 10:00am - 11:00am
Preetha Ramachandran	Sun 5:00pm - 7:00pm
Aasha Reddy	Mon 7:00pm - 9:00pm

Office hours for TAs will begin next week, all times listed are in Eastern Time, Zoom links can be found in Sakai

There is a TA specifically for R support; I'll post his information in Slack.

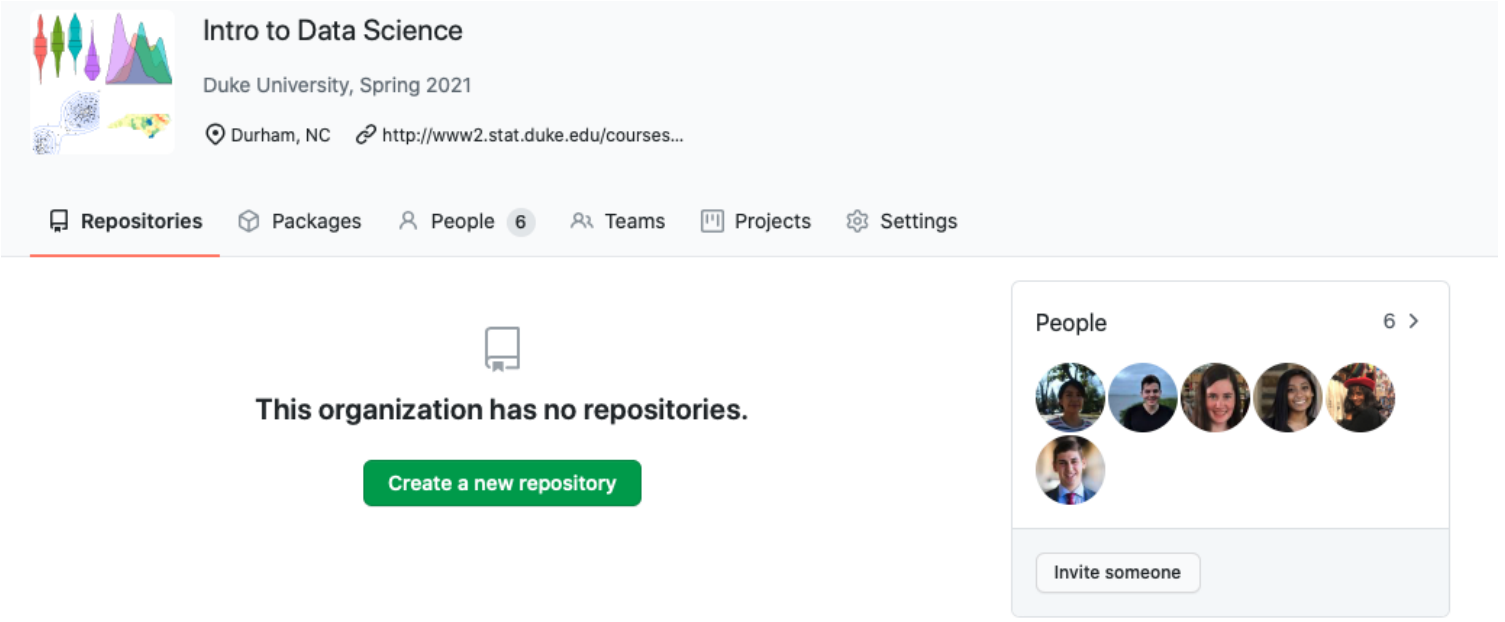
Where to find information

Course website: https://www2.stat.duke.edu/courses/Spring21/sta199.003/



Where to find information

GitHub organization: https://github.com/sta199-sp21-003



Class meetings and structure

Lecture

- Focus on concepts behind data analysis
- Most lectures will be interactive and include examples and hands-on exercises using R/RStudio
- A template R Markdown file will be made available to you for most lectures

Lab

- Focus on computing using R `tidyverse` syntax
- Apply concepts from lecture to case study scenarios
- Work on labs individually or in teams of 3 - 4
- Designed to be finished in 75 minutes or less.

Lectures and labs will be recorded and made available in Sakai.

Textbooks

- Introductory Statistics with Randomization and Simulation
    - Free PDF available online. Hard copy available for purchase.
    - Assigned readings about statistical content
- OpenIntro Statistics, 4th Edition
    - Free PDF available online. Hard copy available for purchase.
    - Assigned readings about statistical content
- R for Data Science
    - Free online version. Hard copy available for purchase.
    - Assigned readings and resource for R coding using `tidyverse` syntax.
- Occasional external readings will be assigned and posted on the course website

Assessments and activities

- **Homework**: Five individual assignments combining conceptual and computational skills.
- **Labs**: Nine individual and team assignments focusing on computational skills.
- **Exams**: Two individual take-home exams.
- **Final project**: One team project in which you use the data science tools to answer a data-based research question. It is due on April 28, 2021.
- **Notes**: In-class examples and practice problems to make lectures interactive and to provide a better understanding of the `R` functions. Due one week following the lecture date.
- **Statistical experience**: One individual assignment for you to engage with statistics outside of the classroom and reflect on your experience.

Grading

Assessment item	Weight
Homework	25%
Labs	15%
Exam 1	20%
Exam 2	20%
Project	15%
Participation and teamwork	2.5%
Statistical experience	2.5%

Late work and regrade requests

- Homework and labs:
  - 24 hour grace period
  - 20% penalty each additional day late
- Late work will not be accepted for all other assessments.
- Regrade requests must be submitted within one week of when the assignment is returned. Only submit a regrade request if there is an error in the grade calculation or a correct answer was mistakenly marked as incorrect.

Policies - sharing / reusing code

Carefully read each assignment so you know what is permitted and what is not. If you are ever unsure what is allowed, please ask me or one of the TAs.

- Similar reproducible examples exist online that will help you answer many of the questions posed on notes, labs, homework assignments, and exams. Use of these resources is allowed unless it is written explicitly on the assignment.
- You must always cite any code you copy or use as inspiration. Copied code without citation is plagiarism.
- Discussion with other students and groups is always allowed unless it is written explicitly on the assignment.

Getting help

- Attend one of the many office hours provided by the teaching team.
- Post your content and course-related questions in our Slack Workspace. The teaching team will continually monitor Slack and respond quickly. Feel free to use Slack or the Zoom chat during lectures to interact.
- Email me your grade and personal related questions.

Getting started

Create a GitHub account

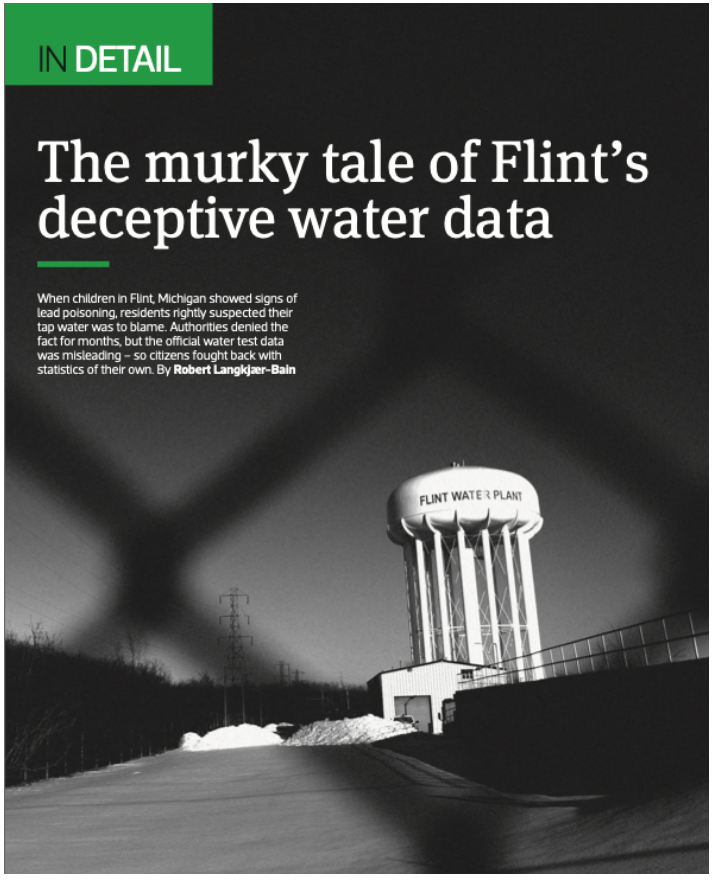
Go to https://github.com/, and create an account (unless you already have one).

Tips for creating a username from Happy Git with R.

- Incorporate your actual name and use all lowercase (not required).
- Pick a username you will be comfortable revealing to your future boss.
- Shorter is better than longer.
- Be as unique as possible in as few characters as possible.
- Make it timeless. Don't highlight your current university, employer, or place of residence.
- Avoid words laden with special meaning in programming, like `NA`.

To-do list before Friday

- Create a GitHub account
- Verify you can access an RStudio Docker Container
- Read *The murky tale of Flint's deceptive water data*
- Skim through a subset of the "Data science in practice" articles on the next slide.



Data science in practice

Take a look at what others have done in data science. Some of these use R, others do not.

- Analyzing trends in the Billboard Hot 100 over the past half century
- Creating interactive redistricting maps
- Tracking their life via Fitbit
- Artificially composing Bach chorales
- Detecting metastatic breast cancer from still images

References

1. Jenny Bryan, Jim Hester. "Happy Git And Github For The User". Happygitwithr.Com, 2021, https://happygitwithr.com/.