Syllabus

INFO 2951 - Introduction to Data Science

Instructor

- Dr. Benjamin Soltoff
- Office: Gates Hall 216
- Email: info2951@cornell.edu
- Office hours: Wednesdays 12-2pm (216 Gates Hall)

Course logistics

- Meets TuTh 10:10am - 11:24am for 28 sessions
- Discussion sections meet Fridays at various times for 15 sessions
- 4 credits, offered for a letter grade
- Prerequisites: ([MATH 1710 or equivalent] AND [CS 1110 or CS 1112]) OR permission of instructor

Course description

This is an applied introductory course for learners who wish to harness growing digital and computational resources. The focus of the course is on using data to identify patterns, evaluate the strength and significance of relationships, and generate predictions using **data**. These techniques are implemented using a **reproducible workflow** through the use of programming languages and version control software. Major emphasis is placed on a pragmatic understanding of core principles of programming and packaged implementations of methods. Students will learn how to use data to make effective arguments, in a way that promotes the ethical usage of data.

Course learning objectives

By the end of the semester, you will…

- Conduct exploratory data analysis through data wrangling and munging as well as visualizations and summary statistics.
- Identify patterns in data to make predictions or to identify associations between variables.
- Evaluate the strength of patterns using statistical and substantive significance.
- Implement data science workflows using common, reproducible methods and software tools.
- Use data ethically and responsibly.

Office hours

- Instructor and undergrad TA OHs - you may attend office hours for any undergrad TA (as well as the instructor!)
- Project mentor OHs - please feel free to meet with your **project mentor** if you have questions about your team projects.

Textbooks

All books are **freely available online**.

R for Data Science, 2e	Grolemund, Wickham	O'Reilly, 2nd edition, 2022
Introduction to Modern Statistics	Çetinkaya-Rundel, Hardin	OpenIntro Inc., 2nd Edition, 2024

Course community

We want you to feel like you belong in this class and are respected. Cornell University (as an institution) and we (as human beings and instructors of this course) are committed to full inclusion in education for all persons. If for any reason you feel that we have failed these goals, please either let us know or report it, and we will address the issue.

Services and reasonable accommodations are available to persons with temporary and permanent disabilities, to students with DACA or undocumented status, to students facing mental health or other personal challenges, and to students with other kinds of learning challenges. Please feel free to let me know if there are circumstances affecting your ability to participate in class. Some resources that might be of use include:

- Office of Student Disability Services: https://sds.cornell.edu
- Cornell Health CAPS (Counseling & Psychological Services): https://health.cornell.edu/services/counseling-psychiatry
- Undocumented/DACA Student support: https://dos.cornell.edu/undocumented-daca-support/undergraduate-admissions-financial-aid

Academic accommodations

We want all students to have the opportunity to be successful in this course. Accommodations can help provide some flexibility and equitable classroom access.

Per university policy, this course provides the following accommodations:

- Disability Accommodations
- Religious-Observance Accommodations
- Title IX Accommodations
- Varsity Athlete Accommodations
- Medical Accommodations
- Military Service
- Other Accommodations

Accessibility

If there is any portion of the course that is not accessible to you due to challenges with technology or the course format, please let me know so we can make appropriate accommodations.

Student Disability Services is available to ensure that students are able to engage with their courses and related assignments. Students should be in touch with Student Disability Services to request or update accommodations under these circumstances.

If you have an approved SDS accommodation, please send a copy of this letter to the instructors at info2951@cornell.edu so we can ensure your accommodations are implemented in this course.

Communication

All lecture notes, assignment instructions, an up-to-date schedule, and other course materials may be found on the course website: info2951.infosci.cornell.edu.

Announcements will be posted through Canvas Announcements periodically. Please check Canvas (or ensure Canvas announcements are forwarded to your email) to ensure you have the latest announcements for the course.

Where to get help

- If you have a question during lecture or discussion, feel free to ask it! There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.
- The course staff is here to help you be successful in the course. You are encouraged to attend office hours to ask questions about the course content and assignments. Many questions are most effectively answered as you discuss them with others, so office hours are a valuable resource. Please use them!
- Outside of class and office hours, any general questions about course content or assignments should be posted on the course discussion forum. There is a chance another student has already asked a similar question, so please check the other posts on GitHub Discussions before adding a new question. If you know the answer to a question posted on the discussion board, I encourage you to respond!

Email

If there is a question that's not appropriate for the public forum, please email us at info2951@cornell.edu. Barring extenuating circumstances, we will respond to INFO 2950 emails within 48 hours Monday - Friday. Response time may be slower for emails sent Friday evening - Sunday.

Activities & Assessment

The activities and assessments in this course are designed to help you successfully achieve the course learning objectives. They are designed to follow the **Prepare, Practice, Perform** format.

- **Prepare**: Includes reading assignments and lectures to introduce new concepts and ensure a basic comprehension of the material. The goal is to help you prepare for the in-class activities during lecture.
- **Practice**: Includes in-class application exercises where you will begin to apply the concepts and methods introduced in the prepare assignment. The activities will graded for completion,

as they are designed for you to gain experience with the statistical and computing techniques before working on graded assignments.

- **Perform**: Includes labs, homework, exams, and the project. These assignments build upon the prepare and practice assignments and are the opportunity for you to demonstrate your understanding of the course material and how it is applied to analyze real-world data.

Lectures (Prepare)

Part of the class time will be lectures that introduce new concepts or review topics from the preparation materials. Lectures will **not** repeat everything in the readings, they will instead highlight important and known to be complex concepts and will be supplemented with live coding activities. You are expected to attend every lecture.

Application exercises (Practice)

A majority of the in-class lectures will be dedicated to working on Application Exercises (AEs). These exercises will give you an opportunity to apply the statistical concepts and code introduced in the prepare assignment. These AEs are due by the end of the day of the corresponding lecture period. Specifically, AEs from Tuesday lectures are due Tuesday by 11:59 pm, and AEs from Thursday lectures are due Thursday by 11:59 pm.

Because these AEs are for practice, they will be graded based on completion, i.e., a good-faith effort has been made in attempting all parts.

The four lowest AE grades will be dropped at the end of the semester.

Labs (Perform)

In labs, you will apply the concepts discussed in lecture to various data analysis scenarios, with a focus on the computation. Most lab assignments will be completed in teams, and all team members are expected to contribute equally to the completion of each assignment. You are expected to use the team's Git repository on the course's GitHub page as the central platform for collaboration. Commits to this repository will be used as a metric of each team member's relative contribution for each lab, and there will be periodic peer evaluation on the team collaboration. Lab assignments will be completed using Quarto, correspond to an appropriate GitHub repository, and submitted for grading in Gradescope.

Labs are due 11:59 pm on the indicated due date.

The lowest lab grade will be dropped at the end of the semester.

Homework (Perform)

In homework, you will apply what you've learned during lecture and lab to complete data analysis tasks. You may discuss homework assignments with other students; however, homework should be completed and submitted individually. Similar to lab assignments, homework must be typed up using Quarto and GitHub and submitted as a PDF in Gradescope.

Homework assignments are due 11:59 pm on the indicated due date.

The lowest homework grade will be dropped at the end of the semester.

Exams (Perform)

There will be two in-person exams during the semester.

- One evening prelim approximately halfway through the course. Date and time TODO by the registrar.
- One final exam.

More details about the content and structure of the exams will be discussed during the semester.

Project (Perform)

The purpose of the project is to apply what you've learned throughout the semester to solve some sort of real-world problem. The project will be completed with your lab teams, and each team will present their work at the end of the semester.

More information about the project will be provided during the semester.

Grading

The final course grade will be calculated as follows:

Category	Percentage
Exams	35%
Prelim	15%
Final exam	20%
Homework	25%
Project	20%
Labs	10%
Application Exercises	10%

The final letter grade will be determined based on the following thresholds:

Letter Grade	Final Course Grade
A+	>= 98
A	93 - 97.99
A-	90 - 92.99
B+	87 - 89.99
B	83 - 86.99
B-	80 - 82.99
C+	77 - 79.99
C	73 - 76.99

Letter Grade	Final Course Grade
C-	70 - 72.99
D+	67 - 69.99
D	63 - 66.99
D-	60 - 62.99
F	< 60

Course policies

Academic honesty

TL;DR: Don't cheat!

Please abide by the following as you work on assignments in this course:

- You may discuss individual homework and lab assignments with other students; however, you may not directly share (or copy) code or write up with other students. For team assignments, you may collaborate freely within your team. You may discuss the assignment with other teams; however, you may not directly share (or copy) code or write up with another team. Unauthorized sharing (or copying) of the code or write up will be considered a violation for all students involved.
- You may not discuss or otherwise work with others on the exams. Unauthorized collaboration or using unauthorized materials will be considered a violation for all students involved. More details will be given closer to the exam date.
- **Reusing code**: Unless explicitly stated otherwise, you may make use of online resources (e.g. StackOverflow) for coding examples on assignments. You may not directly copy and paste from these sources, but instead you need to adapt the code to fit your specific task. You must explicitly cite where you obtained the code using a code comment # immediately near the appearance of the reused code in the file. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- **Use of generative artificial intelligence (GAI)**: Cornell's report on Generative Artificial Intelligence for Education and Pedagogy outlines many of the potential benefits and drawbacks to using GAI in the classroom. In this course, we see the value of coding assistants such as GitHub Copilot and ChatGPT to generate code from text. However as an introductory course, we need to ensure that GAI is not used as a substitute or replacement for student learning. GAI should not be used by students to replace your ability to think clearly. Students who use GAI should use it to **facilitate**, rather than **hinder**, learning.
    - **GAI tools for reference purposes:** You may make use of the technology as a reference tool, similar to looking up the documentation for a function or Googling your problem. For example, I hate writing regular expressions. Absolutely loathe it. Say I have a dataset where I

need to clean a character column to remove all words that are within double asterisk symbols. I might ask ChatGPT

> How do I make a scatterplot using {ggplot2} in R?

- **GAI tools for writing my code/analysis:** You may not make use of the technology to complete substantive portions of your assignments for you. For example, you may not upload your data file to a GAI platform and ask it to create charts and statistical models for you.
- **GAI tools for narrative:** unless instructed otherwise, you may not use GAI to write narrative on assignments. In general, you may use generative AI as a resource as you complete assignments but not to answer the exercises for you.

You are ultimately responsible for the work you turn in; it should reflect *your* understanding of the course content.

TODO update as needed

Any violations in academic honesty standards as outlined in the Cornell University Code of Academic Integrity and those specific to this course will result in a 0 for the assignment (or possibly more) and will be reported to the College of Engineering Academic Integrity Hearing Board.

Extra credit

Students can earn up to a maximum of 1 percentage point towards their final grade through the extra credit assignment. This is the only opportunity for extra credit in the course.

Late work & extensions

The due dates for assignments are there to help you keep up with the course material and to ensure the course staff can provide feedback within a timely manner. We understand that things come up periodically that could make it difficult to submit an assignment by the deadline. Note that the lowest homework and lab assignment will be dropped to accommodate such circumstances.

Late work

- A **slip day** allows you to submit an assignment 24 hours after the deadline and still receive credit without a late penalty. You are provided with a total of **6 slip days** for the entire semester. Slip days may be used on **homework and lab assignments**. You can use up to 1 slip day for a given homework or lab assignment.

  To use your slip days, just submit your assignment late. No need to email telling us you are submitting using your slip days. Check Canvas to see how many of your slip days you have used before submitting an assignment late.

  If you use a slip day, **do not submit anything to Gradescope before the submission deadline**. We may begin grading before the slip day deadline and we will grade whatever submission we see in Gradescope.

If you run out of slip days or fail to submit your assignment prior to the slip day deadline without prior permission then your assignment will not be accepted.

- There is no late work accepted for application exercises, since these are designed to help you prepare for labs and homework.
- There is no late work accepted for project components.
- The late work policy for exams will be provided with the exam instructions.

Waiver for extenuating circumstances

If you need a bit of extra time, **please use your slip days**. Slip days are specifically intended for legitimate reasons for needing an extension like disability, religious observance, Title IX, student athletics, medical problems, and military service.

If using your slip days for accommodations is not working for you or if you have an SDS accommodation which includes deadline flexibility, you may request a deadline extension in-advance of the deadline. We will work with you to develop reasonable accommodations that align with your individual situation.

To request a deadline extension:

1. Commit and push the work you have completed up to this point on the assignment.
2. Email info2951@cornell.edu. In your email clearly state
    a. The assignment
    b. What you have already completed on the assignment.
    c. What you have left to complete.
    d. Your proposed deadline extension (e.g. *Monday, February 8th at 11:59pm.*)

Regrade requests

Regrade requests can be submitted beginning at noon the day after an assignment's grade is posted, and must be submitted on Gradescope within a week of when an assignment is returned. Regrade requests will be considered if there was an error in the grade calculation or if you feel a correct answer was mistakenly marked as incorrect. Requests to dispute the number of points deducted for an incorrect response will not be considered. Note that by submitting a regrade request, the entire question will be graded which could potentially result in losing points.