DS002 Intro to Data Science

updated 1/26

Class	DS002: Intro to Data Science
Instructor	Douglas Goodwin Visiting Assistant Professor in Media Studies, Scripps College Lang 228
Contact	dgoodwin@scrippscollege.edu
Class hours	MW 11:00–12:15
Office Hours	MW 1:00–3:00pm
Discord	https://discord.gg/gtMVkAMV
Textbook	**Data Science from Scratch, 2nd Edition** Joel Grus. ISBN: 9781492041139
Classroom	Zoom, Steele 229
Description	This course is the second part of a two-semester introduction to computer programming and data science. Students will explore, using Python and other tools the nuances of gathering, visualizing and analyzing data to gain insight and intuition with data. Students will be introduced to various data manipulation/analysis, and statistical methods by building their own code from scratch. They will also consider the ethical implications and limitations of creating models of data.

Course Goals

Students completing this course will be able to:

- Define and explain key concepts in the data science
- Learn how to analyze real-world data.
- Gain fluency in basic programming skills in Python with a focus on statistical modeling and machine learning.
- Develop and use essential python programming skills in data analysis, data visualization, and machine learning

We begin by reviewing Python and establishing repositories on GitHub. After review and setup you will read two chapters from “Data Science from Scratch” (DSfS) each week and reproduce the code in the book to build your own code libraries. You are encouraged to write your own code (and tests) as long as you use the same function names. Your code won’t b e efficient, but it will work and your APIs will match dedicated libraries such as *NumPy* (Numerical Python), *pandas* (Python Data Analysis Library), and *scikit-learn* (machine learning in Python).

Please complete the readings before and start your code before class on Monday. I will lecture on the chapters then we will go through your code together during Monday class.

Wednesdays are lab days: work alone or in small groups to use your code to complete the lab assignments.

Points

Activity	Points
10 Weekly Assignments	30 points
1 Presentation	20 points
10 in-class exercises	30 points
Project	30 points
TOTAL	110 points

I give you a little extra room in case you miss an exercise or two.

Points and letter grades

This class will be easy if you keep up with the readings and come to class.

Point range	Grade
90–110	A
80–90	B
70–80	C
60–70	D
< 60	...

Weekly assignments

You will implement the code in the chapters from Data Science from Scratch (DSfS) to build a library of code to use in Deepnote. Push your assignments to your GitHub repository by Monday morning before class. We will use import your code and use it to complete in-class exercises with Deepnote.

Please DON'T copy and paste code! Typing it will help you get familiar with and synthesize the code.

Note the liberal use of the Python's `assert` statement in the sources files. A clean import will give you some assurance that your code is usable. There are more involved ways to test code, even a style of programming called TDD (Test-Driven Development). `assert` statements sprinkled throughout your source code give you TDD-lite!

In-class exercises

Use your code library to solve Data Science problems related to each week's theme. I am creating these exercises now–please ask if you have an idea for an exercise!

Presentation

Sign up for a presentation here: https://forms.gle/AyEUcQ5yJRXJWvCe6

Use this form to select a date and topic to make a 5–10 minute presentation to the class. You may work alone or in teams of less than 4.

We cover two chapters each week, but you only need to choose one topic. Example: on week 02/27 you may choose either Statistics or Probability.

Each chapter contains "Crows" : pointers to topics that are relevant to the subject but not covered in the chapter. These are excellent presentation prompts. Each chapter also concludes with a section called "For Further Exploration" and this is another good jumping off point.

Example: on 02/27 you might tell us why we will want to use SciPy's statistical functions instead of writing our own. What advantages does SciPy offer: speed? convenience? interoperability? all three? Then you might show examples in a Deepnote notebook.

Projects

Complete a small, real-world project in the final week of the class. You can start with one of the in-class exercises or dream up a project of your own. Projects should use external data and be executed in Deepnote or on Google Colab.

Accessibility

If you have a documented disability (physical or cognitive) that may impair your ability to complete assignments or otherwise participate in the course and satisfy course criteria, please meet with us at your earliest convenience to identify, discuss, and document any feasible instructional modifications or accommodations. You should also contact the Accessible Education Office to request an official letter outlining authorized accommodations.

Credits

I will lean heavily on the content in DSfS. Some of the material in this course is based on other classes. We have also heavily drawn on materials and examples found online and tried our best to give credit by linking to the original source. Please contact us if you find materials where the credit is missing or that you would rather have removed.

Weeks, updated 1/26

1. W01 01/17

2. NO CLASS

3. CH01 Introduction

4. W02 01/24

5. CH01 Introduction
6. CH01+ Introduction: GitHub and Deepnote
    i. Class Exercise: Use your GitHub code in Deepnote to visualize the class composition by school

7. W02 01/31

8. CH02 A Crash Course in Python, pt1
9. CH02 A Crash Course in Python, pt2
    i. Class Exercise: Use your GitHub code in Deepnote

10. W03 02/07

11. Python Assignment DUE BEFORE CLASS
12. CH03 Visualizing Data
    i. Video: Data Viz, Computerphile
13. CH04 Linear Algebra
    i. Exercise

14. W04 02/14

15. Viz Assignment DUE BEFORE CLASS

16. CH05 Statistics
17. CH06 Probability
    i. Coinflip Class Exercise, Monty Hall Problem

18. W05 02/21

19. Pandas Assignment DUE BEFORE CLASS

20. CH07 Hypothesis and Inference
21. CH08 Gradient Descent

   i. Class Exercise

22. W06 02/28

23. Assignment DUE BEFORE CLASS

24. CH09 Getting Data
25. CH10 Working With Data
    i. Class Exercise

26. W07 03/07

27. Getting data Assignment DUE BEFORE CLASS

28. CH11 Machine Learning
29. CH12 k-Nearest Neighbors
    i. Class Exercise

30. W09 03/14 SPRING BREAK

31. W08 03/21

32. KNN Assignment DUE BEFORE CLASS
33. CH13 Naive Bayes
    i. Video: Bayes Theorem
34. CH14 Simple Linear Regression
    i. Video: Simple Linear Regression Formula
    ii. Class Exercise

35. W10 03/28
    i. Lineasr Regression Assignment DUE BEFORE CLASS
    ii. CH15 Multiple Regression
    iii. Video: Multiple Regression
    iv. CH16 Logistic Regression
    v. Video: Data Regression
    vi. Class Exercise

36. W11 04/04
    i. Regression Assignment DUE BEFORE CLASS
    ii. CH17 Decision Trees
    iii. CH18 Neural Networks
    iv. Class Exercise

37. W12 04/11
    i. Neural Networks Assignment DUE BEFORE CLASS
    ii. CH19 [Deep Learning]
    iii. CH20 Clustering
    iv. Class Exercise

38. W13 04/18
    i. Clustering Assignment DUE BEFORE CLASS
    ii. CH21 Natural Language Processing
    iii. CH22 Network Analysis
    iv. Class Exercise

39. W14 04/25
    i. NLP Assignment DUE BEFORE CLASS
    ii. CH23 Recommender Systems
    iii. CH24 Databases and SQL
    iv. Class Exercise

40. W16 05/02

    i. PROJECTS

    ii. PROJECTS

    iii. Share on Discord