DSCI511:Data Acquisition and Pre-Processing

Fall 2023
Course Syllabus
College of Computing and Informatics
Drexel University, Philadelphia, PA 19104, USA

Instructor Contact

Dr. Shadi Rezapour

- Email: sr3563@drexel.edu; Office: 1120
- website: https://www.shadirezapour.com/
- Office Hours: Tuesday 2:00 - 3:00 PM, CCI #1120

TA: Ximing Wen

- Email: xw384@drexel.edu
- Office Hours: Mondays 12:00 - 1:00 and Thursday 12:00 -1:00

Email is the best way to contact. Be sure to include a subject line and start the subject with a reference to the course. For example: "DSCI 511 Question" would work. Email without a clear subject may be deleted with spam. When sending emails, make sure you include both the instructor and the course TA.

Student Learning Information

Course Description

Introduces the breadth of data science through a project lifecycle perspective. Covers early-stage data-life cycle activities in depth for the development and dissemination of datasets. Provides technical experience with data harvesting, acquisition, pre-processing, and curation. Concludes with an open-ended term project where students explore data availability, scale, variability, and reliability.

College/Department: College of Computing & Informatics
Repeat Status: Not repeatable for credit
Restrictions: None
Prerequisites: CS 570 (pass the test to waive)

Course Purpose within a Program of Study

This course provides a high-level view of what data is, where it comes from, and how it is used to render insight and support technical products. After introducing the breadth of steps that data scientists take throughout the lifecycle of a project, hands-on and in-depth experience is provided with several early-stage steps that interact closely, and often require iteration before later steps can be approached. This course is a core course in the Data Science Master's program.

Statement of Expected Learning

The course objectives are to:

- obtain an overview of what data is, where it comes from, and what data science entails;
- understand the range of activities essential to a data science project's lifecycle;
- apply technical methods in data collection, construction, and curation; and
- execute a data collection/curation project exploring lifecycle stage interdependence and iteration.

As learning outcomes, students completing this course should be able to use their understanding of project lifecycles, data sources and availability, acquisition and harvesting, pre-processing, and data management to curate data sets of high value that are in alignment with downstream project goals.

Course Materials

Lecture Notes

Students will be responsible for reading these primary course materials in depth, on their own. These lecture notes consist of a collection of interactive Jupyter notebooks, which may be found on the course blackboard website. The course lecture notes are broken down into the following topics, which correspond to a week of content each. More information on course scheduling can be found in that section, below.

Chapter 0: System configuration and processing fundamentals

Chapter 1: Introduction, process, and getting started with data

Chapter 2: Data types and structures: different data, different challenges

Chapter 3: Established collections: databases, dumps, and APIs

Chapter 4: Pre-processing considerations: foresight for downstream needs

Chapter 5: Harvesting content from the World Wide Web

Chapter 6: Data integration and enrichment

Chapter 7: Building and maintaining a robust acquisition stream

Chapter 8: Establishing a database with documentation

Chapter 9: Distribution, accessibility, and data sharing

Chapter 10: Scaling for big data processing

Video Recordings

Each week, I will release a short, pre-recorded video lecture that covers the topics of the week. Please watch these videos prior to each class. These recordings will be distributed through Blackboard.

Weekly exercise

There will be weekly coding exercises released on Bb Learn.

Office Hours

Office hours will be in a question-and-answer format in which students are required to bring any of their questions related to the course content, assignments, and course projects. Before attending office hours, students are expected to send an email to the instructor/TA with the list of their questions.

Texts, Readings, and Resources

Note: All text readings are supplemental to the course lecture notes and will be assigned on a weekly basis. Specific text information is as follows:

- Python Data Science Handbook (PDSH). ISBN: 978-1491912058, O'Reilly Media, 2016.
- Data Science from Scratch (DSFS): First Principles with Python. ISBN: 978-1492041139, O'Reilly Media, 2019
- The Data Science Handbook (TDSH). ISBN:9781119092940, John Wiley & Sons, 2017

Python Data Science Handbook (PDSH) is available free of charge on Github:

- https://jakevdp.github.io/PythonDataScienceHandbook/

and The Data Science Handbook (TDSH) and Data Science from Scratch (DSFS) are available to Drexel students through the University Libraries.

- https://drexel.primo.exlibrisgroup.com/permalink/01DRXU_INST/1pvv3q/alma991021065902604721
- https://drexel.primo.exlibrisgroup.com/permalink/01DRXU_INST/1pvv3q/alma991008049759704721

Required and Supplemental Materials and Technologies

Note: Instructions and discussion of the following materials and technologies are provided in Chapter 0 of the course lecture notes (below). Students are expected to have the following by the start of the first week:

- A GitHub account: https://github.com
- A command line environment with Python (version 3) installed
- The Jupyter notebooks interactive development environment

Assignments, Assessments, and Evaluations

Graded Assignments and Learning Activities

In-Class Quizzes:

We have weekly preparation assessments each session. Each quiz will be counted as a small percentage of your overall grade. They are primarily intended for students as a preparation self-assessment for class sessions. The quiz will be taken in the first 10-15 minutes of each class, with no late/replacement policy.

Homework:

Four structured, individual assignments:

1. Data science programming
2. APIs and pre-processing
3. Web scraping and data integration
4. Disseminating a data processing tool

These assignments will be composed in a modular fashion, with each module/problem worth about 35–45 points apiece (subject to change). Each module can be completed and submitted separately. Modules may be submitted optionally in groups of up to four, consisting strictly of classmates from the same section. Critically important assignment submission information is as follows:

- Submissions will only be accepted through Blackboard (no email submissions)
- For each module, submit only your module-X.ipynb file (Jupyter notebook)
- All relevant submissions will be marked after their Answer Keys have been released
- Any grading disputes must be resolved within 5 days of receipt
- All team member names (even if the submission is individual) must be present in the designated header area in all module notebooks, subject to a 2-point non-compliance penalty.
- All team member names (even if the submission is individual) must be present in the Blackboard submission comments, subject to a 2-point non-compliance penalty.
- All group submissions must be submitted as identical, by each group member, subject to a 2-point non-compliance penalty.
- If you receive help for any question in any individual module, indicate the person's name in both the submission comments and within, subject to a 2-point non-compliance penalty.

Note: assignments will have additional requirements on the structure and format of particular parts of models. For a given part of a module worth, e.g., 'points = 5', these might be:

a. Function(value_portion): programming component in which code must be within the assignment's specification and placed within the specified region of a function. These are _directly_ for credit and worth value_portion * points.

b. Sanity Check: These components are read-only and not for credit, but will be necessary to review in the context of upstream components to complete.

c. Inline(value_portion): worth value_portion * points will have specified output options, e.g., multiple choice, which must be determined from a review of the completed code or a Sanity Check.

Project:

One open-ended group assignment with two phases:

1. Proposal for Data Set Construction and Potential for Use
2. Implementation of Data Set Construction with Documentation and Dissemination

The Project Proposal will be graded based on the following rubric:

a. Forming a team (include in the project proposal team's members self-identified skills, and individual contributions)

b. Discussion of the data source of interest

c. Discussion of dataset potential users and applications

d. Discussion of the plan for acquiring the data

e. Presenting a sample of dataset (it might include preliminary coding)

f. Submission of the project proposal

The Project Implementation will be graded based on the following rubric:

a. Project results, including a final dataset and code which makes it possible to re-construct the dataset

b. Documentation (including Jupyter notebooks, ReadMe file, data dictionary and other supporting documents)

c. Distribution plan

d. Presentation (must have one slide or paragraph of workload distribution of each member in the team)

e. Submission

Teams for Assignment and Project

The assignments and projects are a team effort.

- Each team can have up to 4 members for projects. Students are allowed to work alone in a one-person group.
- Members can always change for each assignment.
- Once a project proposal is submitted, it is recommended to not change members in the project group. If team members change, a written request must be submitted to the instructor.
- Each team submits a single completed project proposal and a single project zip file for the final project.
- Each team will be asked to present the project at the end of the term.

Team Member Project Evaluation:

- All students on a single team initially receive the same grade. However, each team member will evaluate the performance of every other member of his/her/their team. The instructor reserves the right to adjust the grade based on the team evaluation results.
- All students of a team must have their names listed clearly on the front page of the reports. Each team's work must be unique – one team cannot collaborate with another team.

Extra Credits

"The instructor of the course may provide various opportunities for students to earn extra credit through additional homework assignments, research projects, and submissions for the final evaluation, if applicable. These opportunities will be announced during the term, and those interested can take advantage of them to improve their grades. The extra credit is optional.

Grading Matrix

Students will not receive letter grades for individual assignments. Grades are calculated as:

Project: 30% (10% Proposal, 20% Implementation)
Homework: 60%
Quizzes and Class Participation: 10%
Total: 100%

Grade Scale

The following scale will be used to convert points to letter grades:

Note that the instructor may revise this conversion if/when necessary.

Grade	Score	Grade	Score	Grade	Score	Grade	Score
A+	97-100	B+	87-89.9	C+	77-79.9	D+	67-69.9
A	94-96.9	B	84-86.9	C	74-76.9	D	60-66.9
A-	90-93.9	B-	80-83.9	C-	70-73.9	F	0-59.9

Course Schedule

The course's schedule (subject to change, if required) follows the lecture notes at one week per chapter, with the expectation that students will configure systems and review or work through the processing fundamentals in Chapter 0. The regularly scheduled final exam period (to be determined) is reserved for final project presentations. Please observe the following (tentative) schedule and be aware that it may change depending on the term's pace. **Note: all bolded activities are required and must be completed along the tentative timeline.**

	Required Readings	**Supplemental Readings**	**Project**	**Homework**

Week 1	**LN: Chapter 0**
	LN: Chapter 1
		DSFS: 1–2
		TDSH: 1–2, 3.1–3.2
			Group formation;
			begin Phase 1
				Begin Assignment Group 1

Week 2	**LN:Chapter 2**	DSFS: Chapter 9 (All, except "Scraping the Web" and "Using APIs" sections) TDSH: Ch.12 PDSH: Chapters 2.01–2.02	Continue Phase 1	Continue Assignment Group 1
Week 3	**LN: Chapter 3**	DSFS: Chapter 9 ("Using APIs" section)	Continue Phase 1	**Assn. Group 1 Due** **Begin Assn. Group 2**
Week 4	**LN: Chapter 4**	DSFS: Chapter 10 ("Cleaning and Munging", "Manipulating Data", and "Rescaling" sections) TDSH: Chapter 4 (All, except 4.5)	**Begin Phase 2** **Phase 1 Report Due**	Continue Assignment Group 2
Week 5	**LN: Chapter 5**	DSFS: Chapter 9 ("Scraping the Web" section)	Continue Phase 2	**Assn. Group 2 Due** **Begin Assn. Group 3**
Week 6	**LN: Chapter 6**	PDSH: 3.01–3.04, 3.06–3.08	Continue Phase 2	Continue Assn. Group 3
Week 7	**LN: Chapter 7**	None	Continue Phase 2	**Assn. Group 3 Due** **Begin Assn. Group 4**
Week 8	**LN: Chapter 8**	DSFS: Chapter 24 (Supplementary)	Continue Phase 2	Continue Assn. Group 4
Week 9	**LN: Chapter 9**	DSFS: Chapter 27	Continue Phase 2	**Assn. Group 4 Due**
Week 10	**LN: Chapter 10**	None	**Project Phase 2 Due** **Project presentations**	None

Table 1: **LN:** Lecture Notes, **TDSH:** The Data Science Handbook, **PDSH:** Python Data Science Handbook, **DSFS:** Data Science from Scratch

Academic Policies

University Policies

- Academic Integrity, Plagiarism, Dishonesty, and Cheating Policy: http://www.drexel.edu/provost/policies/academic_dishonesty.asp

- Students with Disabilities: Students requesting accommodations due to a disability at Drexel University need to request a current Accommodations Verification Letter (AVL) in the ClockWork database before accommodations can be. made. AVLs are issued by the Office of Disability Services (ODS). These requests are received by Disability Resources (DR), who then issues the AVL to the appropriate contacts. For additional information, visit the DR website at https://drexel.edu/oed/disabilityResources/overview/, or contact DR for more information by phone at 215.895.1401 (V), 215.895.2299 (TTY), or by email at disability@drexel.edu or www.drexel.edu/ods, 3201 Arch St., Street, Suite 210, Philadelphia, PA 19104.

- Course Add/Drop Policy: http://www.drexel.edu/provost/policies/course-add-drop
- Course Withdrawal Policy: http://drexel.edu/provost/policies/course-withdrawal
- Intellectual property: https://drexel.edu/provost/policies-calendars/policies/intellectual_property/

Class Policies

- **The use of artificial intelligence (AI) e.g., ChatGPT, to produce writing as well as codes for this course is not allowed unless it is otherwise stated by the instructor.**
- Class attendance is expected. Attendance includes arriving when the class is scheduled to start and staying for the duration of the class period. Roll may be taken in this course, and you are expected to acknowledge your attendance at least 90% of the time that attendance is taken in the course. I reserve the right to lower any earned course grade if a student fails to meet attendance requirements.
- Late/missed exams/assignments policy: In principle, no late exams/projects will be accepted without prior written approval of the instructor.
- Incomplete policy: Incomplete grades are contingent upon instructor approval and will only be considered in extenuating circumstances beyond the student's control. The instructor is under no obligation to offer an incomplete grade. At least 80% of the graded coursework must have already been completed in order for an incomplete grade to be considered (per the recommendation of the Provost's Office). An incomplete contract with an instructor-determined due date for delivery of the completed work must be completed by the student and the instructor. It can be found here: http://www.drexel.edu/provost/policies/pdf/forms/incomplete.pdf
- The instructor(s) may, at his/her/their discretion, change any part of the course before or during the term, including assignments, grade breakdowns, due dates, and schedule. Such changes will be communicated to students via either email or blackboard announcements.

Communication

- Class syllabus, schedule, readings, assignments, etc. are available on Blackboard and will be updated throughout the term. Always visit Blackboard for updated information.

  The most certain way of getting a message to the instructor or TA is by email. Be sure to put the course number, *DSCI511*, in the subject line. Also, be sure your name is in the message. We get a lot of email messages; putting the course number in the subject line helps us give high priority to students.

  All students are expected to monitor their Drexel email address on a regular basis. You are responsible for making sure that your email accounts are functioning properly.

Class Cancellation

- On rare occasions, instructors may be delayed or unable to attend a scheduled class due to unforeseen circumstances. In the event that an instructor does not appear in class and has not notified the class of his/her expected arrival time, class is cancelled 15 minutes after the scheduled start of class. More information about class cancellations can be found at: https://drexel.edu/provost/policies/cancellation_instructor_absence/

Notice: Appropriate Use of Course Materials

- It is important to recognize that some or all of the course materials provided to you may be the intellectual property of Drexel University, the course instructor, or others. Use of this intellectual property is governed by Drexel University policies, including the policy found here: https://drexel.edu/it/about/policies/policies/01-Acceptable-Use/

  Briefly, this policy states that course materials, including recordings, provided by the course instructor may not be copied, reproduced, distributed or re-posted. Doing so may be considered a breach of this policy and will be investigated and addressed as possible academic dishonesty, among other potential violations. Improper use of such materials may also constitute a violation of the University's Code of Conduct found here: https://drexel.edu/cpo/policies/cpo-1/ and will be investigated as such.