DCDS 510: Data Wrangling

Instructors: Joshua Landman and Amanda Kube

Spring 2021

E-mail: {landman, amanda.kube}@wustl.edu
Office Hours: By Appointment

Location: Zoom
Class Meetings: Tues. 1-4 pm

Course Description

Students studying applications of machine learning to problems arising in the social sciences must learn how to analyze data coming from a variety of sources. Moreover, such data may be flawed in a variety of ways, including missing or erroneous entries, redundant entries, or bias stemming from how the data was collected. This course is intended to provide an introduction to conducting research using methods from data science. We examine the following sources of data: static data sets, dynamic data accessed via an API (Application Programming Interface), web scraping as needed to harvest data where web pages contain desired data but no API is present, and surveys as needed to obtain data not already available. Students ingest data, perform analyses, and document their findings in a scientific and reproducible way, which allows the interspersing of data, code, analysis, and prose, serving to document and make reproducible the scientific process carried out by the students. Students present their results to their peers and faculty with critical reviews of their presentations. Enrollment is limited to students in the DCDS program.

Course Structure

This course's main purpose is to train participating students in data collection methods and writing clear, scientific documentation such that any documented work is able to be easily reproduced by an interested and capable reader. The course will cover topics such as dealing with dirty or flawed data, where to get data and possible sources of data, documenting data processing and analytic pipelines, and more as detailed in the calendar below. These topics will be taught and reinforced through lectures, guest speakers, in-class activities, short-term assignments, and a longer-term project. This course is not intended to be difficult. If you put effort into all required activities and thoughtfully participate in all class meetings, you will probably get an 'A.'

Course Policies

Course Materials and Announcements

We will use Piazza for all questions and discussions related to the class. Please post questions on Piazza rather than sending us email. This serves multiple purposes. First, others may have the same question. Posting to Piazza allows us to clarify the issue for everyone at the same time. Second, we are much more likely to respond in a timely manner if you post on Piazza! Piazza also allows anonymous posting and private posts to instructors. The link to the Piazza discussion for our course can be found here: piazza.com/wustl/spring2021/dcds510

All announcements related to the class will be made in class, on Canvas, or on Piazza. We will assume that any announcement made on Canvas or on Piazza is known to everyone in class within one business day of it being posted. It is important to check Piazza and Canvas regularly! You are responsible for all announcements made in lecture or online.

Course Assignments

There will be 5 short-term assignments and one longer-term final project for the course. We will allow you to complete those assignments in either R or Python depending on your personal preference. However, *at least one* short-term assignment must be completed using each language. You may not do all assignments in R or all assignments in Python. If you do not complete at least one assignment using R and one using Python, *you will not get credit for your final assignment!*

Grading Policy

Your course grade will be calculated as follows:

- Attendance & in-class activities - 15%
- Short-term assignments - 40%
- Final project - 45%

Please submit all assignments on time! Several of these assignemnets will be used in subsequent in-class activities. In addition, you will want feedback on your writing before completing the next assignments. For this reason, late assignments will suffer a 1 point penalty for each day they are late. However, we welcome you to talk to us or to ask for extensions if you are having issues, especially those related to the pandemic.

Topics to be Covered

Topics we will discuss throughout the semester will include but are not limited to:

- Dirty or flawed data
- Where to find static data
- Using an API
- Scraping data

- Collecting new data (survey design)
- Data management
- Data visualization
- Reproducibility
- Scientific writing

Course Calendar

A more detailed calendar will be posted to Piazza and updated as needed. This calendar is subject to change.

Date	Topic	Due
January 26	Dirty Data	Assignment 0
February 2	Exploring Data	Assignment 1
February 9	Exploring Data	
February 16	Fantastic Data and Where to Find Them	Assignment 2
February 23	Fantastic Data and Where to Find Them	
March 2	**Wellness Day - No Class**	
March 9	Fantastic Data and Where to Find Them	
March 16	Collecting New Data	Assignment 3
March 23	Collecting New Data	
March 30	Maintaining Data & Final Project	Assignment 4
April 6	Maintaining Data	
April 13	Maintaining Data	
April 20	TBA	Assignment 5
April 27	Final Project	
May 4	Final Project	
May 11	**Reading Week - No Class**	Final Project

Accommodations Due to COVID-19

Due to the pandemic and to the nature of the course, all class meetings and faculty presentations will be held over Zoom. Attendance and participation are still expected and to this end, we ask that you keep your video on so we can assess attendance. Microphones can be muted during presentations, but not during class discussions. We will continue to adjust any COVID-19 policies as we get further information from the administration or as the situation progresses over time, so please be flexible as we are all navigating this together.

Other Accommodations

Accommodations based upon sexual assault: The University is committed to offering reasonable academic accommodations (for example, no contact order, course changes) to students who are victims of relationship or sexual violence, regardless of whether they seek criminal or disciplinary action. If you need to request such accommodations, please contact the Relationship and Sexual

Violence Prevention Center (rsvpcenter@wustl.edu or 314-935-3445) to schedule an appointment with an RSVP confidential, licensed counselor. Information shared with counselors is confidential. However, requests for accommodations will be coordinated with the appropriate University administrators and faculty. See http://rsvpcenter.wustl.edu. If a student comes to us to discuss or disclose an instance of sexual assault, sex discrimination, sexual harassment, dating violence, domestic violence or stalking, or if we otherwise observe or become aware of such an allegation, we will keep the information as private as we can, but as faculty members of Washington University, we are required to immediately report it to a Department Chair or Dean or directly to Ms. Jessica Kennedy, the University's Title IX Coordinator. If you would like to speak with the Title IX Coordinator directly, Ms. Kennedy can be reached at (314) 935-3118, jwkennedy@wustl.edu, or by visiting the Title IX office in Umrath Hall. Additionally, you can report incidents or complaints to the Office of Student Conduct and Community Standards, or by contacting WUPD at (314) 935-5555 or your local law enforcement agency. See https://titleix.wustl.edu/.

Bias Reporting: The University has a process through which students, faculty, staff and community members who have experienced or witnessed incidents of bias, prejudice or discrimination against a student can report their experiences to the University's Bias Report and Support System (BRSS) team. See: http://brss.wustl.edu Mental Health Mental Health Services' professional staff members work with students to resolve personal and interpersonal difficulties, many of which can affect the academic experience. These include conflicts with or worry about friends or family, concerns about eating or drinking patterns, and feelings of anxiety and depression. See: http://shs.wustl.edu/MentalHealth

Preferred Name and Gender Inclusive Pronouns: In order to affirm each person's gender identity and lived experiences, it is important that we check in with others about pronouns. This simple effort can make a profound difference in a person's experience of safety, respect, and support. See: https://students.wustl.edu/gender-pronouns-information/ and https://registrar.wustl.edu/student-records/ssn-name-changes/preferred-name/

Military Service Leave: Washington University recognizes that students serving in the US Armed Forces and their family members may encounter situations where military service forces them to withdraw from a course of study, sometimes with little notice. Students may contact the Office of Military and Veteran Services at (314) 935-2609 or veterans@wustl.edu and their academic dean for guidance and assistance. See: https://veterans.wustl.edu/policies/ policy-for-military-students/.

Center for Diversity and Inclusion (CDI): The Center of Diversity and Inclusion (CDI) supports and advocates for undergraduate, graduate, and professional school students from underrepresented and/or marginalized populations, creates collaborative partnerships with campus and community partners, and promotes dialogue and social change. One of the CDI's strategic priorities is to cultivate and foster a supportive campus climate for students of all backgrounds, cultures and identities. See: diversityinclusion.wustl.edu/