Information Exploration

INFO 3401; Fall 2020

Monday, Wednesday, Friday; 10:20–11:10
CASE W262

Abram Handler
Instructor, Information Science
E-mail: abe.handler@colorado.edu
Office: https://cuboulder.zoom.us/my/abehander
Office hours: Tuesdays, 14:00–15:00

Brian Keegan, Ph.D.
Assistant Professor, Information Science
E-mail: brian.keegan@colorado.edu
Office: https://cuboulder.zoom.us/my/brianckeegan
Office hours: Fridays, 11:30–13:30

Course Description

The volume, velocity, variety, and variability of data challenges our ability to collect, analyze, interpret, and act. This course will develop students' skills and sensibilities for conducting *exploratory data analysis* across seven kinds of data. Students will learn to collect, analyze, visualize, evaluate, and communicate data to motivate new questions, make predictions, and work towards solutions. This course will call upon the quantitative and computational skills students have developed in previous courses and will increase their confidence and autonomy as data analysts and scientists who can deliver insights from diverse kinds of data.

Learning objectives

- Improve students' confidence analyzing diverse kinds of data
- Develop students' ability to match questions to data to solutions
- Understand professional data science tools and methods
- Think critically about the opportunities and limitations of data

Course Design

Class will meet three times per week (Monday, Wednesday, Friday) from 10:20–11:10 on Zoom with the possibility of an optional on-campus component in CASE W262 after September 25. We will be employing a "flipped classroom" in which students watch pre-recorded lectures before class and class time prioritizes discussions, coding exercises, and sharing work-in-progress. Student performance will be evaluated through Module Assignments, Module Quizzes, Participation, and a Final Project. There is no final exam.

The class is split up into seven two-week modules organized around different types of data: (1) tabular, (2) relational, (3) temporal, (4) spatial, (5) dyadic, (6) structured, and (7) unstructured data. The first week of each module will focus on the fundamentals of the data type and the second week of each module will explore applications of the data type.

Requirements and contingencies for COVID-19
As a matter of public health and safety due to the pandemic, all members of the CU Boulder community and all visitors to campus must follow university, department and building requirements, and public health orders in place to reduce the risk of spreading infectious disease. Required safety measures at CU Boulder relevant to the classroom setting include:

1. maintain 6-foot distancing when possible
2. wear a face covering in public indoor spaces and outdoors while on campus consistent with state and county health orders
3. clean local work area
4. practice hand hygiene
5. follow public health orders
6. if sick and you live off campus, do not come onto campus (unless instructed by a CU Healthcare professional), or if you live on-campus, please alert CU Boulder Medical Services

Students who fail to adhere to these requirements will be asked to leave class, and students who do not leave class when asked or who refuse to comply with these requirements will be referred to Student Conduct and Conflict Resolution. See the policies on COVID-19 Health and Safety and classroom behavior and the Student Code of Conduct. If you require accommodation because a disability prevents you from fulfilling these safety measures, please see the "Accommodation for Disabilities" statement on this syllabus.

Before returning to campus, all students must complete the COVID-19 Student Health and Expectations Course. Before coming on to campus each day, all students are required to complete a Daily Health Form. Students who have tested positive for COVID-19, have symptoms of COVID-19, or have had close contact with someone who has tested positive for or had symptoms of COVID-19 must stay home and complete the Health Questionnaire and Illness Reporting Form remotely. In this class, if you are sick or quarantined, please email us if you will require accommodations for any illness (including COVID-19) such as extensions and incompletes.

- **Conditionally hybrid**. The class will meet entirely remotely for weeks 1–5 (August 24–September 25) and there will be no on-campus lectures or office hours. During the first week, the instructors and students will have a discussion about whether, when, and how to shift to a "hybrid" instructional model combining remote and on-campus instruction. During the (remote) lecture on September 25 (the Friday before week 6) the instructors will announce their determination whether the class will shift to a "hybrid" format agreed upon by the instructors. We will make this determination based on the case data reported by the Boulder County Department of Public Health. Specifically, if there are more than 12 new cases reported per day[1] for more than 7 of the previous 14 days, then the subsequent week will remain remote.
- **Flipped classroom**. We will employ a "flipped classroom" design, which means an approximately 50 minute lecture will be pre-recorded and should be viewed on Canvas before class. We will use our valuable class time for more active learning experiences like exercises, demonstrations, discussions, and sharing work-in-progress. *Attendance is required.* Students will be expected to have watched the lectures before starting class, attendance will be taken for each class, and there will be regular quizzes to assess student comprehension of the pre-class lectures.
- **Synchronous instruction**. We will use a "synchronous remote" instructional model: students will attend on Zoom at the scheduled time. *Attendance is required.* In order to fully participate remotely, students will need access to a personal computer with a web camera and a reliable high-speed internet connection. We will also experiment with using breakout rooms, whiteboards, annotations, polling, and other Zoom features throughout the semester. We expect students to be professional and engaged while on Zoom. Please arrive on time, ensure your mic and video are working beforehand, use the "Raise Hand" functionality to request to be unmuted to ask or answer questions, dress appropriately and have a background appropriate for a university class, and refrain from side activities and conversations in or outside of Zoom. Students

[1] Fewer than 4 new cases per 100,000 people is a popular criterion for reopening. Boulder County has 300,000 residents.

who violate these norms will be warned once and then removed and lose the day's participation credit if they persist in behavior that disrupts the class. We will discuss, debug, and develop practices about how best to run this instructional model and are open to feedback and iteration throughout the term.

- **Illness**. Should a student contract any illness that requires mandatory sequestration, intensive medical treatment, or extended convalescence and disrupts their ability to participate in class and complete assignments, the instructors will try to accommodate their condition without penalty with extensions and incompletes. This also applies if the student has a family member whose diagnosis, treatment, and recovery will affect their ability to participate. *Please do not ghost us*: students should notify the instructors as soon as possible of events that will impact their engagement with the class so that we can triage and develop an accommodation plan rather than scrambling at the end of the semester.

Prerequisites
Students should have completed the sequences of INFO 2201 and INFO 2301 or similar coursework covering intermediate computational reasoning and intermediate statistical reasoning before enrolling in INFO 3401. If you have questions, please email Dr. Handler or Dr. Keegan.

Course Website and Materials
There is no textbook required for class, but there will be required readings, tutorials, and other material, which will be made available through Canvas:

Canvas: https://canvas.colorado.edu/courses/62560
Zoom: https://cuboulder.zoom.us/j/98569655581

Once the semester begins, this PDF version of the syllabus will be revised infrequently and any revised requirements will be posted as announcements and updated course schedule to Canvas. The instructors reserve the right to make changes to the course's schedule, evaluation criteria, policies, *etc.* through announcements in class and on Canvas, so please check Canvas regularly. If you have questions, please email Dr. Handler or Dr. Keegan.

Computing Requirements
Students will need to use statistical computing software as well as teleconferencing software to participate in class. Jupyter notebooks written in Python 3 will be used for all in-class examples and assignments. The Anaconda distribution of Python 3.5 (or above) is *strongly* recommended to provide all of these programs and other libraries. Lectures will include exercises and presentations with the expectation that students participate with their own computers. If students do not have access to a computer to use for computing or Zoom, they should immediately email Dr. Handler or Dr. Keegan to work out an alternative arrangement. Students are welcome to use an alternative *programmatic* (*not* Excel or Tableau) data analysis environment like R, Matlab, Julia, *etc*., but instructional support will only be provided for Anaconda and Python. Students who require technical assistance should email the instructors with the code and data they are working with, a summary of their debugging efforts to date, and attend an instructor's office hours.

Evaluation
Students will be evaluated through four different mechanisms.

- **Module Assignments** (40%). Module Assignments are intended to develop students' confidence and skill conducting their own exploratory data analyses. There are eight Module Assignments in total, one per module. Each Module Assignment is due by 10am the Wednesday after the start of the next module. The format and evaluation criteria of each Module Assignment will vary. In the absence of an approved excuse, late submissions will lose an additional 2% of their grade for every hour elapsed since the deadline: *assignments submitted after Friday at 12:00 (50 hours after the deadline) will lose all credit.* The lowest module assignment grade will be automatically dropped.

- **Module Quizzes** (20%). Module Quizzes are intended to assess students' progress in learning the fundamental data analysis skills from the lectures and readings posted to Canvas. There will be at least one quiz per module, which may occur as either pre-scheduled or unannounced ("pop quiz") events at the instructors' discretion. In the absence of an approved excuse, missed quizzes cannot be made up but the lowest quiz grade will be automatically dropped.
- **Participation** (20%). Participation will be assessed on a combination of attendance and engagement. The "flipped classroom" model will have much more active engagement than a traditional lecture involving discussions, case studies, and other "hands-on-keyboard" activities. Students should expect to be "cold called" to present (using Zoom's screen share functionality) during class time. If a student has a disability, anxiety, or another issue that limits their ability to participate in this format, please email Dr. Handler or Dr. Keegan. Participation grades cannot be made up but students can miss up to five classes without an effect on their grade. Students who will have extended absences due to a medical condiction, injury, or family emergency should contact the instructors as soon as possible to develop an accommodation plan.
- **Final Project** (20%). The Final Project is intended to be a portfolio piece highlighting a student's data collection, analysis, and visualization abilities. The project will be both a data analysis and write-up with the goal of submitting for external publication as a guest blog post, op-ed, *etc*. A proposal will be presented in the middle of the term to go through the critical response process (2% of final grade), will be presented to the class in the final week (2% of final grade), and a written version will be posted to the class's Medium publication (16% of final grade). Further details about the Final Project will be collaboratively developed and detailed later in the course. In the absence of an approved excuse, late Final Project submissions will be docked 2% of their value for every hour elapsed since the deadline.

Course Policies

In-Class Confidentiality

The success of this class depends on students feeling comfortable sharing questions, ideas, concerns, and confusions about assignments, work-in-progress, and their personal experiences. Students may read, comment, and run on classmates' writing, code, and other class-related content for the sole purpose of use within this class. However, students may not use, run, copy, perform, display, distribute, modify, translate, or create derivative works of another student's work outside of this class without that student's expressed written consent or formal license. Furthermore, students may not create any audio, video, or other records during class time without the instructor's permission nor may students publicly share comments made in class attributable to another person's identity without that person's permission.

Instructor Interaction

Dr. Handler and Dr. Keegan will check e-mail between 8:00 and 18:00 on non-holiday business days and try to respond to emails within 24 hours. They welcome online or offline interactions outside of class, however these are not appropriate spaces for discussing class matters. E-mailing Dr. Handler or Dr. Keegan or coming to our (remote) office hours are the best ways to get help and feedback outside of lecture.

Accommodations for Disabilities

We are committed to providing everyone the support and services needed to participate in this course. If you qualify for accommodations because of a disability, please submit your accommodation letter from Disability Services to the instructor in a timely manner so that your needs can be addressed. Disability Services determines accommodations based on documented disabilities in the academic environment. Information on requesting accommodations is located on the www.colorado.edu/disabilityservices/students. Contact Disability Services at 303-492-8671 or dsinfo@colorado.edu for further assistance. If you have a temporary medical condition or injury, see Temporary Medical Conditions under the Students tab on the Disability Services website and discuss your needs with the instructors.

Religious Observance
Campus policy regarding religious observances requires that faculty make every effort to deal reasonably and fairly with all students who, because of religious obligations, have conflicts with scheduled exams, assignments or required assignments/attendance. If this applies to you, please e-mail Dr. Handler or Dr. Keegan as soon as possible to make the appropriate accommodations.

Classroom Behavior
Students and instructors each have responsibility for maintaining an appropriate learning environment. Those who fail to adhere to such behavioral standards may be subject to discipline. Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, color, culture, religion, creed, politics, veteran's status, sexual orientation, gender, gender identity and gender expression, age, ability, and nationality. Class rosters are provided to the instructor with the student's legal name. The instructor will honor your request to address you by an alternate name or gender pronoun. Please advise the instructors of this preference early in the semester so that we may make appropriate changes. For more information, see the policies on class behavior and the student code.

Harassment and Discrimination
The University of Colorado Boulder (CU Boulder) is committed to maintaining a positive learning, working, and living environment. CU Boulder will not tolerate acts of sexual misconduct, discrimination, harassment or related retaliation against or by any employee or student. CU's Sexual Misconduct Policy prohibits sexual assault, sexual exploitation, sexual harassment, intimate partner abuse (dating or domestic violence), stalking or related retaliation. CU Boulder's Discrimination and Harassment Policy prohibits discrimination, harassment or related retaliation based on race, color, national origin, sex, pregnancy, age, disability, creed, religion, sexual orientation, gender identity, gender expression, veteran status, political affiliation or political philosophy. Individuals who believe they have been subject to misconduct under either policy should contact the Office of Institutional Equity and Compliance (OIEC) at 303-492-2127. Information about the OIEC, the above referenced policies, and the campus resources available to assist individuals regarding sexual misconduct, discrimination, harassment or related retaliation can be found at the OIEC website.

Honor Code
All students enrolled in a University of Colorado Boulder course are responsible for knowing and adhering to the academic integrity policy of the institution. Violations of the policy may include: plagiarism, cheating, fabrication, lying, bribery, threat, unauthorized access to academic materials, clicker fraud, resubmission, and aiding academic dishonesty. All incidents of academic misconduct will be reported to the Honor Code Council (honor@colorado.edu; 303-735-2273). Students who are found responsible for violating the academic integrity policy will be subject to nonacademic sanctions from the Honor Code Council as well as academic sanctions from the faculty member. Additional information can be found at honorcode.colorado.edu.

Acknowledgements

This syllabus was typeset in LaTeX using Overleaf with the fbb/Bembo font and is derived from the `memoir` styles adapted by Kieran Healy and Benjamin 'Mako' Hill.

Course Outline

The schedule will evolve throughout the semester, so please consult the schedule online at Canvas for the most up-to-date information.

Week 1 – Introductions

Monday, August 24; Wednesday, August 26; Friday, August 28

Course overview; configuring environment; building a data scientist mindset.

Week 2 – Tabular data: Fundamentals

Monday, August 31; Wednesday, September 2; Friday, September 4

Reviewing tabular data structures, file I/O; long *vs.* wide data; tabular data with `pandas`.

Week 3 – Tabular data: Applications

Wednesday, September 9; Friday, September 11

(No class on Monday, September 7 in observance of Labor Day)

Cleaning and analyzing data about populations from U.S. Census, United Nations, WHO, *etc.*

Week 4 – Relational data: Fundamentals

Monday, September 14; Wednesday, September 16; Friday, September 18

Understanding related data across multiple tables; types of joins; databases with `sqlite`.

Week 5 – Relational data: Applications

Monday, September 21; Wednesday, September 23; Friday, September 25

Administering a simple database; analyzing historical baseball data from Retrosheet and Lahman Database.

Week 6 – Temporal data: Fundamentals

Monday, September 28; Wednesday, September 30; Friday, October 2

Understanding time series data; auto-correlation, change-points, anomalies; forecasting using `prophet`.

Week 7 – Temporal data: Applications

Monday, October 5; Wednesday, October 7; Friday, October 9

Preparing and forecasting economic data from FRED, World Bank, IMF, *etc.*

Week 8 – Spatial data: Fundamentals

Monday, October 12; Wednesday, October 14; Friday, October 16

Understanding spatial data; projections, choropleths, spatial joins, lookups; visualization using `geopandas`.

Week 9 – Spatial data: Applications

Monday, October 19; Wednesday, October 21; Friday, October 23

Analyzing and visualizing political data from DataVerse, ICPSR, MIT Election Lab, *etc.*

Week 10 – Dyadic data: Fundamentals

Monday, October 26; Wednesday, October 28; Friday, October 30

Understanding networks; network types, data structure, metrics, and dynamics; analysis using `networkx`.

Week 11 – Dyadic data: Applications

Monday, November 2; Wednesday, November 4; Friday, November 6

Preparing and analyzing network data from ICON, KONECT, Network Repository, *etc*.

Week 12 – Structured data: Fundamentals

Monday, November 9; Wednesday, November 11; Friday, November 13

Parsing and navigating structured data like XML and JSON; scraping using `requests` and `BeautifulSoup`.

Week 13 – Structured data: Applications

Monday, November 16; Wednesday, November 18; Friday, November 20

Scraping, storing, and analyzing structured data from APIs like Wikipedia, Twitter, Spotify, *etc*.

Week 14 – Unstructured data: Fundamentals

Monday, November 23; Wednesday, November 25

(No Fall Break this year, we will be class the Monday and Wednesday before Thanksgiving.)

Analysis of unstructured data; text pre-processing, vectorizing, metrics; analysis using `nltk`.

Weeks 15 – 16: Unstructured data – Applications

Monday, November 30; Wednesday, December 2; Friday, December 4; Monday, December 7

(Last class will be Monday, December 7 and there is no final exam.)

Pre-processing and analyzing text data from literature, journalism, and social web.

Module	Week	Dates	Topics
	1	Aug 24 – Aug 28	Introductions, data science mindset
Tabular	2	Aug 31 – Sep 4	**Fundamentals**: single tables, `pandas`
	3	Sep 9 – Sep 11	**Applications**: population data
Relational	4	Sep 14 – Sep 18	**Fundamentals**: multiple tables, `sqlite`
	5	Sep 21 – Sep 25	**Applications**: sports data
Temporal	6	Sep 28 – Oct 2	**Fundamentals**: time series, `prophet`
	7	Oct 5 – Oct 9	**Applications**: economic data
Spatial	8	Oct 12 – Oct 16	**Fundamentals**: mapping, `geopandas`
	9	Oct 19 – Oct 23	**Applications**: political data
Dyadic	10	Oct 26 – Oct 30	**Fundamentals**: network analysis, `networkx`
	11	Nov 2 – Nov 6	**Applications**: social network data
Structured	12	Nov 9 – Nov 13	**Fundamentals**: JSON & XML, `BeautifulSoup`
	13	Nov 16 – Nov 20	**Applications**: API data
Unstructured	14	Nov 23 – Nov 27	**Fundamentals**: text processing, `nltk`
	15 – 16	Nov 30 – Dec 7	**Applications**: text data