CDS DS 100: Data Speaks Louder than Words

Instructor Name: TBA Course Dates: TBA

Office Location: TBA Course Time & Location: TBA

Contact Information: TBA Course Credits: 4

Office Hours: TBA

TA/TF/Learning Assistant information, if relevant: TBD

TF Name: Graham Albert

Office Location: TBA

Contact Information: graham@bu.edu

Office Hours: M/W 9:00am-10:30am

Course Description: This course introduces students to three perspectives that are fundamental to their ability to reason with data: critical thinking, inferential thinking, and computational thinking. All three of these perspectives are integral to the data-driven research processes that are common in data science, thus allowing students to learn and practice how they may be able to make hypotheses, test these hypotheses, and construct or deconstruct arguments that are rooted in data.

The course starts with the use of data to model and hence understand real-world phenomena through the use of public data sets (both curated or scraped) with a particular focus on data reflecting societally-relevant themes (e.g., public health, education, and environment). The course focuses on the use of model summarization, data visualization, and model-based simulations to interpret and communicate our understanding of these real-world phenomena as well as to the potential for bringing these derived models to bear on real-world questions and applications (e.g., comparing different policies).

Particular emphasis is placed on exposing students to and developing their appreciation for the principles underlying data mining and machine learning methods, including regression, classification and clustering, and the statistical concepts of measurement error and prediction. The course teaches critical concepts and skills in computer (Python) programming, linear regression, and statistical inference. It also delves into dilemmas surrounding data analysis such as balancing individual privacy and social utility.

Hub Learning Outcomes

Social Inquiry I (SO1)

- **Learning Outcome #1**: Students will identify and apply major concepts used in the social sciences to explain individual and collective human behavior including, for example, the workings of social groups, institutions, networks, and the role of the individual in them.

Students will employ hands-on analysis of real-world datasets, including curated economic data, data scraped from digital collections, social networks, etc. In that context, the course will expose students to social and legal issues surrounding data analysis, including issues of privacy and data ownership, and will highlight the many ways in which data could be used (or misused).

Students will be looking at data from multiple vantage points . For example, by looking at data characterizing Covid-19 infections, hospitalization, death, vaccination, students would be able to differentiate between phenomena (e.g., correlations) identified at the macro scale (federal and state) versus those identified at the micro scale (cities and communities) and to draw conclusions or make statements supported with evidence from data (e.g., impact of socioeconomic background).

Students will be encouraged to apply what they learn on societally-relevant case studies of their choice (e.g., case studies similar to those presented in https://www.callingbullshit.org/) by applying the tools and techniques covered in class to analyze data sets in order to support or debunk hypotheses.

Digital/Multimedia Expression (DME):

- **Learning Outcome #1**: Students will be able to craft and deliver responsible, considered, and well-structured arguments using media and modes of expression appropriate to the situation.

Students are asked to use real data to understand relationships and patterns while also being introduced to critical concepts and skills in computer programming and statistical inference. In order to build their arguments, students will use multimodal data analysis and visualization in ways that are appropriate to the task at hand. This would include:

- Generation and interpretation of scatter plots, histograms, bar charts, box plots,
- Making predictions using simple regression
- Characterizing data quality and communicating associated uncertainties
- Establishing confidence in the reproducibility of predictions
- Reaching defensible conclusions about real-world questions

These skills will be taught and evaluated in both problem sets and laboratory exercises, as well as in exams (e.g., asking students to critique statements made in light of a specific visualization, or asking them to select from a set of suggested visualizations the one that would either support or deconstruct an argument).

- **Learning Outcome #2**: Students will be able to demonstrate an understanding of the capabilities of various communication technologies and be able to use these

As part of DS 100, students will be introduced to multiple forms of data visualization and presentation, for example including histograms, scatterplots, word clouds, heat maps, infographics, etc. Each one of these forms of communication can be particularly effective (or even misleading) in certain settings. For example, the choice of different scales (e.g., absolute vs relative change) on an axis could over or under-emphasize particular conclusions from the data.

Given the multitude of sources from which the data is collected, students would be exposed to proper ways of handling the data. For example, to preserve the privacy of individuals or communities in a large data set, students will be introduced to the use of randomization techniques (blurring the data). As another example, to deal with the scale of data it may be necessary to only consider/analyze a subset of all observations. In that context, students will be introduced to various ways in which selection bias may influence conclusions they may be able to reach with implications on reproducibility.

- **Learning Outcome #3**: Students will be able to demonstrate an understanding of the fundamentals of visual communication, such as principles governing design, time-based and interactive media, and the audio-visual representation of qualitative and quantitative data.

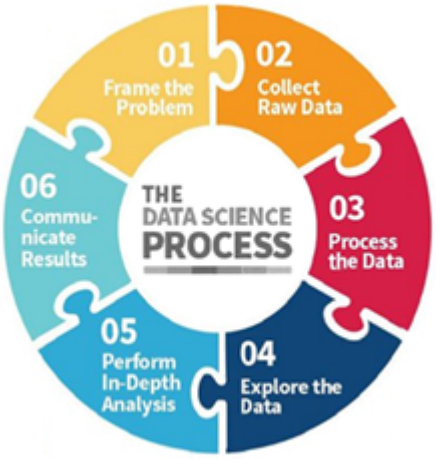
Students will learn how to use Python to organize and manipulate data in tables, and to visualize data effectively. Students will use computation to help their data tell a story through fundamental principles and methods of data visualization. The data used throughout this course would include longitudinal data (time series over long-time scales), geospatial data (data overlaid on apps), or both. These modalities offer students different ways to interact with the data. For example, with time series data, students will be able to develop animations to show how phenomena or inferences may evolve over time. As another example, with geospatial data sets, students will be able to develop animations or heatmaps that may project different messages/narratives based on the level of aggregation (e.g., achieved by zooming in and out).

In all of the above learning outcomes, we note that some student work products will be in the form of multimedia reports, in which data visualization is coupled with narratives or video clips. For example, in a report on deforestation due to climate change, students may add audio or video clips to demonstrate change over time. Students may also include their own narration to supplement and/or add texture to the graphs, heatmaps, etc.

Research and Information Literacy (RIL) Learning Outcomes

The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. In discussion sections (worth 10% of the grade), students will work in small teams, working under the supervision of the teaching fellow to frame a question or test a hypothesis using a set of potential data sources. The key phases of that process are the exploration and identification of relevant data sets,

the formulation and reformulation of the questions based on the identified data, the development of a set of data processing/analytics steps leading to an answer, and the interpretation and/or validation of the answer. To a large extent, going through these phases mirrors the six steps of the data science research process shown below.



This course emphasizes learning through doing: students will work on large real-world data sets through interactive assignments to apply the skills they learn. Throughout, the underlying thread is that data science is a way of thinking, not just an assortment of methods. Students will hone their interpretation and communication skills, which are essential skills for data scientists. Students will explore the proper way to complete the research process, eliminating bias (including their own).

As such, students will be trained in the processes underlying robust inference from real-world data from a variety of domains, run experiments and test their hypotheses, know the correct statistical tools to use depending on the task, quantify and understand uncertainty in data, and understand and utilize computation and simulation in data science. Students will learn to articulate the benefits and limits of computing technology for analyzing data and answering questions.

- **Learning Outcome #1:** Students will be able to search for, select, and use a range of publicly available and discipline-specific information sources ethically and strategically to address research questions.

In addition to publicly-available datasets, which will be made available to them, students will also be encouraged to proactively identify data sets available online which may be relevant to the question at hand – e.g., to refine findings or to compare findings across populations – possibly leading students to rephrase the question. In that respect, steps #2 and #4 will be exemplified through use cases

covered in lectures, problems that students have to work on as part of homework assignments, and student group activities pursued in discussion sections.

As part of their framing of questions in step #1 and their selection and exploration of data sets in steps #2 and #4, students will be introduced to and encouraged to think through the social issues surrounding data processing and analysis in steps #3 and #5, such as transparency, privacy, and inclusive design. As with the identification and exploration of data and information sources, the consideration of ethical dimensions will be covered in the use cases used in lectures and discussion sections, with students expected to reflect on them and to apply them in write-ups of their reports in which they communicate results in step #6.

- **Learning Outcome #2:** Students will demonstrate understanding of the overall research process and its component parts, and be able to formulate good research questions or hypotheses, gather and analyze information, and critique, interpret, and communicate findings.

Students will learn the underpinnings of the overall data science research process by repeatedly covering its key phases (of problem framing, data identification and exploration, data processing and analysis, and interpretation and communication of the results). To that end, the course uses a spiral approach, in which students iterate over these phases (steps #1 to #6), with each iteration being more involved by virtue of students' use of new approaches or consideration of additional data sets. For example, while a first iteration may consider descriptive statistics, a second iteration may consider building confidence around key statistics, a third may involve correlative analysis or regression, and a fourth may involve hypothesis testing using new data sets.

While different homework assignments and projects will focus on one or more of the phases of the data science research process (shown in the figure above), the entire process will also be covered explicitly in one of the early lectures and/or discussion sections in the course (lecture hours #1 and #2), and will be reinforced throughout the semester through the iterative application of different (and increasingly sophisticated) methods. In doing so, students will develop an understanding of the entire process in addition to learning (and being assessed on) the specific research methods involved in each one of the phases. Finally, the course will end with a "tour de force" of going through the entire data science research process in the last "putting it all together" part of the course (lecture hours #39 and #40 and associated project #3).

One of the themes emphasized throughout the course is the importance of reproducible conclusions. Students may be given a data set to support a hypothesis, and then are given another data set to try and prove the same hypothesis. For example, students will go through the exercise of checking each other's conclusions, as data conclusions should be reproducible by the different students using the same data or by the same student given two data sets. As another example, in a homework assignment or in a discussion section group activity, students

will be given a question and a proposed analysis of a data set, and asked to explain whether the analysis addresses the question and how the analysis could change and still address the question. Through these specific drills students will be able to develop their ability to not only interpret and communicate their own findings, but also evaluate and critique research findings by others.

Assessment of RIL Learning Outcomes:

Assessment of and feedback on RIL learning outcomes will be provided to students as part of the regular evaluation of homework assignments and discussion section activities.

The following are specific examples that illustrate how students will be assessed on their choices of data sets on the one hand, and their choices of research methods on the other.

- In homework assignments (e.g., homework #2 and #3) or programming project assignments (e.g., project #3) asking students to apply a particular research method in order to answer a question (e.g., using correlation to expose racial disparities in COVID infections), students will be given choices of data sets and will have to justify (and will be graded on) their choice of the data sets (or subsets) they use to develop their arguments.
- In homework assignments (e.g., homeworks #5 through 9) or programming project assignments (e.g., project #2 and #3) asking students to choose a method out of many to apply to a given data set (e.g., using clustering versus regression, or using selection followed by a join or vice versa), students will have to discuss and elaborate on the societal or ethical implications of their choice (with a distinct percentage of the grading rubric allocated to that aspect).

For grading purposes, students will be informed of the rubric/schemas used to assess their work -- what percentages of their grade correspond to choices made that are related to choices made and justification provided for the research methods/process (e.g., proper consideration of social and ethical implications in a final report, or veracity of the analysis based on building confidence intervals, or on documentation for reproducibility, etc.)

Similar assessment will be done for group activities pursued in discussion sections, in which students would be divided into a small number of small groups and asked to make and justify choices of datasets and/or research methods directed at answering a specific question or testing a specific hypothesis. This format provides an opportunity for peer evaluation (with one group justifying their choices while another critiquing it). Here the assessment is less focused on "grading the choices" and more focused on "active participation in the drill."

Additional Learning Objectives of Course:

TBD

Books and Other Course Materials

Inferential Thinking: by Ani Adhikari and John DeNero, with contributions by David Wagner and Henry Milner.

An optional resource the instructor may make use of to identify case studies for students to analyze is https://www.callingbullshit.org/. The students will not be required to purchase any accompanying materials.

The language used to teach the course is Python, and is supplemented using a library called datascience. The source code can be found at the attached link here.

Courseware

The datascience Package: The datascience package is an open source Python package that helps make programming more accessible to all students, regardless of background. As a pedagogical aid, the package is designed to help students more intuitively conduct data science techniques without first spending considerable time directly learning more complex tools such as pandas or matplotlib.

The full documentation to the datascience package can be found here, but students typically only need the Python Reference Guide for all the functions that are used widely in the course.

This class will use Piazza to post homework assignments, lab activities, and information about the projects, as well as discussion boards and blog posts.

Assignments and Grading

This course will require students to participate in weekly lectures, a required weekly lab, short-term weekly homework assignments, and longer-term (~monthly) projects in which they will tackle real-life issues using real, publicly-available data. Students will also complete a mid-term exam and a final exam.

Certain project assignments (e.g., project #3 on "putting it all together") will require students to make use of multimedia components (audio or visual aids) to supplement data visualizations and student narrations of said data visualizations and their conclusions.

Grades will be assigned using the following weighted components:

Activity	**Grade**
Class and on-line participation	5%
Lab activities and drills	15%
Homework assignments (10)	20%
Programming and data analysis projects (3)	15%
Midterm Exam	15%
Final Exam	30%

Educational strategies to encourage students' full engagement in the course both inside and outside of classroom or activity space:

Active learning is emphasized through interactive lectures to deliver course material and activities and assignments to ensure deep engagement with course material and relevant tools. During lectures, the instructor will pose a variety of questions to students during class, await student responses and/or call on students individually. Active learning will also be explored throughout the labs, homeworks and projects, which give students the opportunity to directly apply the tools covered in lectures and discussion sections to real-world datasets. The projects and case studies allow the students to engage with the methods and challenges of data analysis, including eliminating bias, testing hypotheses, quantifying and understanding uncertainty in data, and understanding and utilizing computation and simulation in data science.

Consistent, comprehensive feedback will be given directly to students who are struggling during class, as well as on in-class and homework assignments. Students will be encouraged to attend office hours for assistance as needed.

Resources/Support/How to Succeed in This Course:

1. To succeed in this course students should come to class having read the material beforehand, attend all lectures, come to discussion prepared with questions, complete all assignments on time, and discuss problems and material with your fellow classmates.
2. Students are welcomed and encouraged to visit office hours.
3. The Education Resource Center offers free individual and group tutoring.
4. Accommodations for Students with Documented Disabilities: If you are a student with a disability or believe you might have a disability that requires accommodations, please contact the Office for Disability Services (ODS) at (617) 353-3658 or access@bu.edu to coordinate any reasonable accommodation requests. ODS is located at 25 Buick Street on the 3rd floor.

Community of Learning: Class and University Policies

1. Course members' responsibility for ensuring a positive learning environment (e.g., participation/ discussion guidelines). If for some reason you are not able to attend a lecture, discussion, or lab session, please let the appropriate instructor know so that the appropriate accommodations may be made. If you know that you will miss a lab, you need to contact the members of your group to have them fill you in on what you missed.

   BU Policy on Religious Observance

2. **Attendance & Absences.** Clearly state your attendance policy, limit on absences, etc., including any implications of class attendance on grading. List all unusual required meetings (e.g., field trips, guest speakers). Affirm Policy on Religious Observance.

   TBD

3. **Assignment Completion & Late Work**. Detail your policy regarding how students should submit assignments (in person, by email, on courseware site, etc.) as well as how you will address late work, missed exams, etc.

TBD

4. **Academic Conduct Statement**

   Students are expected to abide by the guidelines and rules of the Academic Code of Conduct *https://www.bu.edu/academics/policies/academic-conduct-code/*

Outline of Class Meetings: Timeline for topics, readings, labs, and assignment due dates

Note: Reading are from the course textbook at https://inferentialthinking.com/chapters/intro.html

Class Hours	Topics covered in Lectures & in Discussion Sections	Textbook Readings	Homework Assignments, Lab Work, and Projects
1	Introduction to Data Science and to the course: Syllabus and expectations		
2	Data Science Research Process: Data lifecycle and the iterative nature of data-driven research	1.1, 1.2, 1.3	Lab #1: Basic Python programming with variables and expressions
3	Data as the reflection of causes and effects of complex processes	2	Homework #1
4	Representing relationships as tables	3	
5	Data Types and Operations	4, 5	Lab #2: Table Operations
6	Using Python to building and management of Tables	6.1, 6.2	Homework #2
7	Use Case: Census Data	6.3, 6.4	
8	Data Collection, Processing, and Exploration: Data Wrangling	7, 7.1	Lab #3: Creating & Extending Tables
9	Data Exploration, Summarization and Visualization: Charts and Histograms	7.2, 7.3	Homework #3
10	Data Transformation: Functions	8, 8.1	
11	Data Transformations: Groups	8.2, 8.3	Project #1 (covering phases 1-2 of the DS Research Process)
12	Data Transformations: Pivots and Joins	8.4	
13	Table Examples	8.5	Lab #4: Functions & Visualizations
14	Iteration	9, 9.1, 9.2, 9.3	Homework #4
15	Chance	9.5, 18.1	
16	Sampling	10, 10.1, 10.2	Lab #5: Conditional Statements, Iteration Tables
17	Models	10.3, 11.1	Homework #5
18	Comparing Distributions	11.1, 11.2	
19	Decisions and Uncertainty	11.3	Lab #6: Assessing Models
20	A/B Testing	11.4, 12.1, 12.2	Homework #6
21	Causality	12.3	
22	Examples	12.2	Lab: Midterm Review
23	Midterm		
24	Bootstrap	13, 13.1, 13.2	
25	Confidence Intervals	13.3, 13.4	Lab #7: Bootstrap
26	Interpreting Confidence Intervals	14, 14.1, 14.2	Homework #7
27	Center and Spread	14.3, 14.4	
			Project #2 (covering

Class Hours	Topics covered in Lectures & in Discussion Sections	Textbook Readings	Homework Assignments, Lab Work, and Projects
1	Introduction to Data Science and to the course: Syllabus and expectations		
2	Data Science Research Process: Data lifecycle and the iterative nature of data-driven research	1.1, 1.2, 1.3	Lab #1: Basic Python programming with variables and expressions
3	Data as the reflection of causes and effects of complex processes	2	Homework #1
4	Representing relationships as tables	3	
5	Data Types and Operations	4, 5	Lab #2: Table Operations
6	Using Python to building and management of Tables	6.1, 6.2	Homework #2
7	Use Case: Census Data	6.3, 6.4	
8	Data Collection, Processing, and Exploration: Data Wrangling	7, 7.1	Lab #3: Creating & Extending Tables
9	Data Exploration, Summarization and Visualization: Charts and Histograms	7.2, 7.3	Homework #3
10	Data Transformation: Functions	8, 8.1	
11	Data Transformations: Groups	8.2, 8.3	Project #1 (covering phases 1-2 of the DS Research Process)
12	Data Transformations: Pivots and Joins	8.4	
13	Table Examples	8.5	Lab #4: Functions & Visualizations
14	Iteration	9, 9.1, 9.2, 9.3	Homework #4
15	Chance	9.5, 18.1	
16	Sampling	10, 10.1, 10.2	Lab #5: Conditional Statements, Iteration Tables
17	Models	10.3, 11.1	Homework #5
18	Comparing Distributions	11.1, 11.2	
19	Decisions and Uncertainty	11.3	Lab #6: Assessing Models
20	A/B Testing	11.4, 12.1, 12.2	Homework #6
21	Causality	12.3	
22	Examples	12.2	Lab: Midterm Review
23	Midterm		
24	Bootstrap	13, 13.1, 13.2	
25	Confidence Intervals	13.3, 13.4	Lab #7: Bootstrap
26	Interpreting Confidence Intervals	14, 14.1, 14.2	Homework #7
27	Center and Spread	14.3, 14.4	
			Project #2 (covering