DS 100: Data Speak Louder than Words Syllabus

Course

- All course communication should take place on Piazza: https://piazza.com/bu/fall2023/ds100/home
- Class Location: LAW AUD
- Class Time:
    - 2:30-4:45 (14:30-16:45) – typically less (see below)
- Lab Location: CDS 164, KCB 103, SAR 300, CDS 262
- Course Dates: Fall 2023
- Course Credits: 4

Course Description

In this course we will introduce you to three fundamental perspectives for reasoning 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 you to learn and practice how you can make and test hypotheses, and construct or deconstruct arguments that are rooted in data.

We will first use public data sets (both curated or scraped) focused on socially-relevant themes (e.g., public health, education, and environment) to model and understand real-world phenomena. We will focus on using model summarization, data visualization, and model-based simulations to interpret and communicate our understanding of these real-world phenomena as well as the potential for bringing these derived models to bear on real-world questions and applications (e.g., comparing different policies).

Particular emphasis will be placed on exposing you to and developing your 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. We will teach you critical concepts and skills in computer programming (Python), linear regression, and statistical inference. We will also delve 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.*

We will employ hands-on analysis of real-world datasets, including curated economic data, data scraped from digital collections, social networks, and more. In this context, the course will expose you 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).

In this course we will be looking at data from multiple vantage points. For example, by looking at data characterizing COVID-19 infections, hospitalizations, deaths, vaccinations, we will 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 draw conclusions or make statements supported with evidence from data (e.g., impact of socioeconomic background).

We will encourage you to apply what you learn on societally-relevant case studies of your 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.*

We will use real data to understand relationships and patterns while also introducing critical concepts and skills in computer programming and statistical inference. In order to build your arguments, you will use multimodal data analysis and visualization in ways that are appropriate to the task at hand. This will include:

- Generation and interpretation of scatter plots, histograms, bar charts, and box plots
- Making predictions using simple regression
- Characterizing data quality and communicating associated uncertainties
- Establishing confidence in reproducible 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 you to critique statements made in light of a specific visualization, or asking you 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 technologies ethically and effectively.*

As part of DS-100, we will introduce you to multiple forms of data visualization and presentation, 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, you will be exposed to proper ways of handling the data. For example, to preserve the privacy of individuals or communities in a large data set, and 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, we will introduce you to various ways in which selection bias may influence conclusions you may be able to reach with implications on reproducibility.

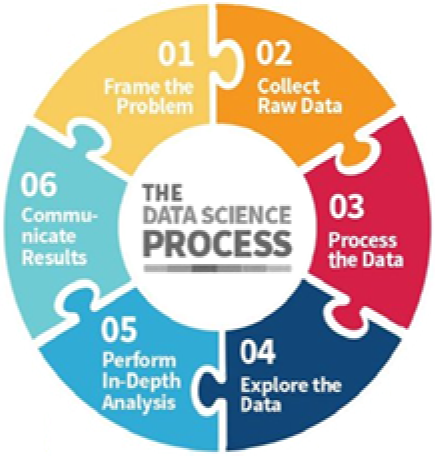
Learning Outcome #3*: Students will be able to demonstrate an understanding of the fundamentals of visual communication, such as principles governing design, time-based and interactive media, and the audio-visual representation of qualitative and quantitative data.*

We will teach you how to use Python to organize and manipulate data in tables, and to visualize data effectively. Furthermore, you will be able to use computation to help your data tell a story through fundamental principles and methods of data visualization. The data used throughout this course will include longitudinal data (time series over long-time scales), geospatial data (data overlaid on apps), or both. These modalities will offer you different ways to interact with the data. For example, with time series data, you will be able to develop animations to show how phenomena or inferences may evolve over time. As another example, with geospatial data sets, you will be able to develop animations or heat maps 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 of your 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, you may add audio or video clips to demonstrate change over time. You may also include your own narration to supplement and/or add texture to the graphs, heatmaps, etc.

Research and Information Literacy (RIL) Learning Outcomes

We will teach you 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 your grade), you 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: we will work on large real-world data sets through interactive assignments to apply the skills you learn. Throughout, the underlying thread is that data science is a way of thinking, not just an assortment of methods. We will focus on honing your interpretation and communication skills, which are essential skills for data scientists. Moreover, we will explore the proper way to complete the research process, eliminating bias (including your own).

As such, you will be trained in the processes underlying robust inference from real-world data from a variety of domains, run experiments and test your 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. You 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, you 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 you to rephrase the question. In that respect, steps #2 and #4 of the data science process will be exemplified through cases covered in lectures, problems that you will have to work on as part of homework assignments, and student group activities pursued in discussion sections. A set of recommended sources can be found on the course website.

As part of your framing of questions in step #1 and your selection and exploration of data sets in steps #2 and #4, you 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. You will be expected to reflect on them and to apply them in write-ups of your reports in which you communicate your 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.*

You will learn the underpinnings of the overall data science research process by repeatedly covering its key phases (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 you will iterate over these phases (steps #1 to #6), with each iteration being more involved by virtue of your 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, and will be reinforced throughout the semester through the iterative application of different (and increasingly sophisticated) methods. In doing so, you 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.

One of the themes emphasized throughout the course is the importance of reproducible conclusions. For example, you may be given a data set to support a hypothesis, and then could be given another data set to try and prove the same hypothesis.

Assessment of Research and Information Literacy (RIL) Learning Outcomes:

Assessment of and feedback on Research and Information Literacy (RIL) learning outcomes will be provided to you as part of the regular evaluation of homework assignments and discussion section activities.

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

- In homework assignments or programming project assignments asking you to apply a particular research method in order to answer a question (e.g., using correlation to expose racial disparities in COVID infections), you will be given choices of data sets and will have to justify (and will be graded on) your choice of the data sets (or subsets) you will use to develop your arguments.
- In homework assignments or programming project assignments asking you 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), you will have to discuss and elaborate on the societal or ethical implications of your choice (with a distinct percentage of the grading rubric allocated to that aspect).

Books and Other Course Materials

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

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

The language used to teach the course is Python, and is supplemented by a number of open source libraries, including one unique to the teaching of Data Science: datascience.

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 you more intuitively conduct data science techniques without first spending considerable time directly learning more complex tools such as numpy, pandas or matplotlib.

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

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

Top Hat: Please hold off on Top Hat until a future post on Piazza.

~~To evaluate your understanding of course concepts over the duration of the semester we will be using Top Hat. Top Hat is a learning management system that will enable the course instructors to determine the class's understanding of a concept in real time and adjust their teaching accordingly. Please use the following link to create an account: https://app.tophat.com/register/ and then type in the six digit sign up code: 133071. Top Hat Basic should suffice for this course.~~

Assignments and Grading

This course will require you to participate in lectures, a required weekly discussion section, short-term weekly homework assignment & labss, and longer-term (~monthly) projects. The projects will tackle real-life issues using real, publicly-available data. You will also complete 2 exams and a final exam.

Certain project assignments will require you to use multimedia components (audio or visual aids) to supplement data visualizations and narrations of said data visualizations and with your conclusions.

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.

Grades will be assigned using the following weighted components:

Activity	**Grade**
Class and on-line participation	10%
Lab activities and drills	15%
Homework assignments (10)	25%
Programming and data analysis projects (2)	20%
Exams	15%
Final Exam	15%

Attendance & Absences

Missing more than three classes may affect your final grade by negatively impacting your participation components. Please note that **attending discussion is mandatory**. You may not attend a lab section that is not your designated lab section without permission from Professor White. If for some reason you are not able to attend a lecture, discussion, exam, or lab session, please let the appropriate instructor via Piazza know so that the appropriate accommodations may be made. Some special cases exist, e.g. BU Policy on Religious Observance.

Assignment Completion & Late Work

Communication, particularly written, is essential to data science. As a result, we expect correct spelling, grammar, naming, etc. All work submitted will be evaluated for communication quality and will impact the score of the work. Remember, you can always ask someone to proofread your work (while honoring the Academic Code of Conduct). This is a good practice to start.

a. In general, homework will be released during the Wednesday lecture and due, via Gradescope, at the start of the following Wednesday lecture. Homework may be up to 24 hours late with a -10% impact on score. Homework will not be accepted after 24 hours late. However, check the schedule at the end of this document for details.

b. The homework may be completed, via Gradescope, by 11:59 PM (23:59h) on the Friday after it is released for extra credit. There are a couple of exceptions to this rule as noted in the schedule below.

c. Labs are assigned during the Monday lecture (generally). Labs are generally expected to be completed during the lab periods following lecture. However, they can be submitted as late as the following Sunday at 11:59 PM (23:59h) for full marks. Labs may be submitted 24 hours late (that is Monday at 11:59 PM/23:59h) with a -10% impact on the score. Labs will not be accepted after this late submission deadline.

d. Projects will generally have 3-4 weeks to be completed. The details for each project will be clearly stated in Piazza/Gradescope. Projects can be up to 48 hours late with a -10% impact on score. Projects will not be accepted after 48 hours late.

Academic Conduct Statement

You are expected to abide by the guidelines and rules of the Academic Code of Conduct. In addition, CDS has developed a Policy on the Use of AI Text Generation (aka ChatGPT) which will generally apply in this class. However, some aspects of the class will have an explicit directive to not leverage AI tools which will be clearly stated and expected to be adhered to.

Integrity & Conduct

We take the Student Responsibilities guide very seriously and in particular: "civility and respect for others within the University." In this class we should all strive to be the model for what we want our University and industry to be.

How to Succeed in this Course

1. In brief: To succeed in this course you 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 materials with your fellow classmates.

2. You are welcomed and **encouraged** to visit office hours.

3. Use Piazza to ask questions about course material. This term we will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from classmates and the instructional team. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. If you have any problems or feedback for the developers, email team@piazza.com. Find our class page at: https://piazza.com/bu/fall2023/ds100.

4. The Education Resource Center offers free individual and group tutoring. We are also hoping to build a community with this class and, as such, would love to support study groups, etc. We encourage DS-100 alumni to participate as lab assistants, tutors, and teaching fellows.

Schedule (tentative)

Note: Readings are expected to be read before the following lecture. Lectures, discussions and a few, non-standard due dates get their own rows in this table. Legend: (D) = Assignment Due

Date	Topics covered in Lectures & in Discussion Sections	Readings	Homework Assignments, Lab Work, and Projects
W. Sep 06	Introduction to Data Science and to the course: Syllabus and expectations	1.1, 1.2, 1.3, 2	
F. Sep 08	Discussion TBD		
M. Sep 11	Data Science Research Process: Data lifecycle and the iterative nature of data-driven research; Cause & Effect	3	
W. Sep 13	Representing relationships as tables	4, 5	HW 01 (Due: Sep 20)
F. Sep 15	Discussion TBD		
M. Sep 18	Data Types and Operations	6.1, 6.2, 6.3, 6.4	Lab 01 (Due: Sep 24)
W. Sep 20	Using Python: building and management of Tables, Use Case: Census Data	7, 7.1	HW 02 (Due: Sep 27); HW 01 (D)
F. Sep 22	Discussion TBD		
M. Sep 25	Data Collection, Processing, and Exploration: Data Wrangling	7.2, 7.3	Lab 02 (Due: Oct 01)
W. Sep 27	Data Exploration, Summarization and Visualization: Charts and Histograms	8, 8.1, 8.2, 8.3	HW 03 (Due: Oct 04); HW 02 (D); Project 1 (Due: Oct 13); HW 02 (D)
F. Sep 29	* Discussion * Project 1 Questions		
M. Oct 02	Data Transformation: Functions; Groups	8.4, 8.5	Lab 03 (Due: Oct 08)
W. Oct 04	Data Transformations: Pivots and Joins; Table Examples	8 (complete)	HW 03 (D)
F. Oct 06	* Discussion * Exam Review * Project 1 Questions		Project 1: Checkpoint (D)

Date	Topics covered in Lectures & in Discussion Sections	Readings	Homework Assignments, Lab Work, and Projects
M. Oct 09	No Class (Holiday)	8 (complete)	
T. Oct 10	* Grouping, Pivots, Joins * Exam #01	9.5, 18.1, 10, 10.1, 10.2	
W. Oct 11	Chance; Sampling	9, 9.1, 9.2, 9.3, 10.3, 11.1	HW 04 (Due: Oct 18)
F. Oct 13	* Discussion * Project 1 Questions		Project 1 (D)
M. Oct 16	Iteration; Models	11.1, 11.2, 11.3	Lab 04 (Due: Oct 22)
W. Oct 18	Comparing Distributions; Decisions and Uncertainty	11.4, 12.1, 12.2	HW 05 (Due: Oct 25); HW 04 (D)
F. Oct 20	Discussion TBD		
M. Oct 23	A/B Testing	12.2, 12.3	Lab 05 (Due: Oct 29)
W. Oct 25	Causality; Examples	13 (complete)	HW 06 (Due: Nov 01); HW 05 (D)
F. Oct 27	* Discussion * Exam 02 Review		
M. Oct 30	Bootstrap; Confidence Intervals	13 (complete)	
W. Nov 01	* Interpreting Confidence Intervals * Exam #02	14.1, 14.2	HW 06 (D)
F. Nov 03	Discussion TBD		
M. Nov 06	Ethics	14.1, 14.2	Project 02 (Due: Dec 01)
W. Nov 08	Ethics	14.3, 14.4	HW 08 (Due: Nov 15)
F. Nov 10	Discussion TBD		
M. Nov 13	Center and Spread	14.6	Lab 06 (Due: Nov 19)
W. Nov 15	The Normal Distribution and Sample Means	14.6	HW 08 (D)
R. Nov 16	Project 02: Checkpoint 01		Project 02: Checkpoint 01 (D)
F. Nov 17	Discussion TBD		

Date	Topics covered in Lectures & in Discussion Sections	Readings	Homework Assignments, Lab Work, and Projects
M. Nov 20	Classification & Classifiers	15 (complete)	Lab 07 (Due: Nov 26)
T. Nov 21	Project 02: Checkpoint 02		Project 02: Checkpoint 02 (D)
W. Nov 22	Thanksgiving Recess		
F. Nov 24	Thanksgiving Recess		
M. Nov 27	Designing Experiments	15 (complete)	Lab 08 (Due: Dec 03)
W. Nov 29	Correlation; Linear Regression		HW 09 (Due: Dec 06)
F. Dec 01	Discussion TBD		Project 02 (D)
M. Dec 04	Least Squares; Residuals	16	
W. Dec 06	Regression Inference	17, 17.1, 17.2, 17.3, 17.4	HW 10 (Due: Dec 11); HW 09 (D)
F. Dec 08	Discussion TBD		
M. Dec 11	Updating Predictions	18	HW 10 (D)
M. Dec 18	*LAW AUD: 3:00pm – 5:00pm (**Tentative**, Final Exam Schedule will be released mid semester typically)*		