# CAS CS 701
## Tools for Data Science
## Fall 2023

**Meeting Place:** CDS 164

**Meeting Time:** TR 9:30 - 10:45am

**Instructor:** Prof. Mark Crovella

- **Office:** CDS 937
- **Office Hours:** T 3-4, Th 4-5.
- **Office Hours Location:** CCDS 9th floor
- **Email:** `crovella@bu.edu`

**Teaching Fellow:** Mr. Gabriel Franco

- **Office Hours:** M 2:30 - 3:30, W 2 - 4.
- **Office Hours Location:** CCDS 9th floor
- **Email:** `gvfranco@bu.edu`.

## Overview of the Course

This course is a Master's level introduction to data science, focusing on proficiency in working with and analyzing data. The course emphasizes practical skills in working with data, while introducing students to a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. The goal of the class is to provide to students a hands-on understanding of classical data analysis techniques and to develop proficiency in applying these techniques in a modern programming language (Python).

Broadly speaking, the course breaks down into three main components, which we will take in order of increasing complication: (a) unsupervised methods; (b) supervised methods; and (c) methods for structured data.

Lectures will present the fundamentals of each technique; focus is not on the theoretical analysis of the methods, but rather on helping students understand the practical settings in which these methods are useful. Class discussion will study use cases and will go over relevant Python packages that will enable the students to perform hands-on experiments with their data.

## Prerequisites

Prerequisites: Students taking this class **must** have prior familiarity with programming, at the level of DS110, CS105, CS108, or CS111, or equivalent. In this course you will use python – you are assumed to either know python, or be ready to learn it quickly on your own in the first week. Linear Algebra – DS121 or CS132 or equivalent (MA 242, MA 442) – is **required.** DS210 or CS112 is also helpful.

## Learning Outcomes

Students who successfully complete this course will be proficient in data acquisition, manipulation, and analysis. They will have good working knowledge of the most commonly used methods of clustering, classification, and regression. They will also understand the efficiency issues and systems issues related to working on very large datasets.

## Textbook and Slides

The textbook used in the course is published at `http://mcrovella.github.io/CS701-Tools-for-Data-Science/`. This online text will evolve as the course progresses, but I will work to keep it up-to-date.

The slides I use in lecture are actually executable Jupyter notebooks. When I show you code in lecture, it will almost always be runnable code in the form of Jupyter notebooks which you can download and execute on your own computer. You can modify them any way you'd like, play around with them, experiment, etc.

The notebooks and everything else I use in lecture are published on `github`. The repository is `https://github.com/mcrovella/CS701-Tools-For-Data-Science`. If you want to clone or fork the repository using `git`, please feel free. If you find a bug, feel free to submit a pull request.

Some of the lectures are based on *Introduction to Data Mining,* by Tan, Steinbach and Kumar. This is a good place to go for more detail if some methodological aspect is not clear. For up-to-date reference on Pandas, scikit-learn, or any of the other software tools we use, there is no substitute for online resources. Google will quickly bring you to the authoritative (and current) references on software tools.

## Tools and Platforms

We will use:

1. Piazza for questions (`https://piazza.com/bu/fall2023/ds701/`),
2. Github for homeworks, midterm and project submission,
3. Gradescope for grading and grade management, and
4. Kaggle for the midterm.

You should already be signed up for Gradescope (if not, enroll using code **JK3ZV8**). You can add yourself to Piazza if you are not already enrolled.

You will need an account on Github. Once you have them, fill out the form at `https://docs.google.com/forms/d/e/1FAIpQLSeWQDpIojbSiiqUs_`

_eSNai9g32SASxnpdZM75064uDOFKOXw/viewform?usp=sf_link to let us know what it is.

If you don't have an up to date Python installation, take care of that right away.

## Piazza

We will be using Piazza for class discussion. The system is really well tuned to getting you help fast and efficiently from classmates, the teaching fellows, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. Our class Piazza page is at: https://piazza.com/bu/fall2023/da701/. We may also use Piazza for distributing materials such as homeworks and solutions.

When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However please *don't* provide answers to homework questions on Piazza. It's OK to tell people *where to look* to get answers, or to correct mistakes; just don't provide actual solutions to homework questions.

## Programming Environment

We will use python as the language for teaching and for assignments that require coding. Instructions for installing and using Python are in the online textbook.

## Course and Grading Administration

Homeworks are due at midnight on the date shown on the syllabus. Assignments will be submitted using github and gradescope. Please review the instructions for submitting homeworks, on the Resources page of Piazza.

*NOTE: IMPORTANT:* Late assignments **WILL NOT** be accepted. However, you may submit **one** homework up to 3 days late. You **must** email Mr. Franco before the deadline if you intend to submit a homework late.

Final grades will be computed based on the following:

**20%** Midterm

**40%** Homework assignments.

**40%** Final Project

The exact cutoffs for final grades will be determined after the class is complete.

## Homeworks

There will four homework assignments. In a typical assignment you will analyze one or more datasets using the tools and techniques presented in class.

Homeworks will be submitted via github. For this, we need your github account (create one if you don't already have it). After you have created it, fill out the form at https://docs.google.

com/forms/d/e/1FAIpQLSeWQDpIojbSiiqUs__eSNai9g32SASxnpdZM75064uDOFKOXw/viewform?usp=sf_link to let us know what it is.

You are expected to work individually on homeworks.

## Midterm

The midterm will be a Kaggle Data Science competition among the students in the class with a live leaderboard. Students will need to submit predictions based on a training dataset and a report detailing the methods used and decisions made. Note that the intent is not to use the leaderboard to determine your grade, but rather to help you assess how effective your work is. Accordingly, 80% of the grade will be based on the report and only 20% will be based on the competition score related to the quality of the predictions made.

## Project

A major goal of this course is to gain experience with real-world data science problems in form of a project. For the project you will extract some knowledge or conclusions from the analysis of dataset of your choice. The analysis will be done using a subset of the methods we described in class.

Grading will be based on specific deliverables as well as your performance in your team throughout the semester.

For the final project, students may get the opportunity to work with BU Spark! on a real world, data-driven project for a company, non-profit, or institution. Spark projects have already been curated and will be presented during "Pitch Day". Project descriptions will be made available at the start of the semester.

Once every student has a final project, every team will need to upload a SCRUM file to the final project repository every week which gives a short report on the status of their project.

SCRUM is an agile method used in many software companies. Fast and concise, it is a short report answering the following questions:

- What have I worked on?
- What will I be working on next?
- Have I run into any issues? Do I need help?
- Have I talked to the client recently? When are we meeting with them next?

### Project Expectations

- All team members should contribute equally and proactively to project work; we will evaluate team contributions through a peer evaluation at the end of the semester and this will be factored into your grade.
- You / your team lead should make yourself available to speak with your client on a bi-weekly basis (depends on client availability)
- You / your team lead should meet with your Spark PM on a weekly basis
- You should meet with your team every other day (can / should be a short meeting)

- For any team communication issues, please let your spark PMs know asap - they are here to help. If the problem persists please email me with a description of the situation.
- All students are expected to abide by University conduct policies as detailed in the following links:
  - Boston University Student Codes of Conduct: `https://www.bu.edu/policies/policy-category/student-codes-condcut/`
  - College of Arts & Sciences Codes of Conduct: `https://www.bu.edu/cas/academics/undergraduate-education/academic-conduct-code/resources-for-students/`
  - Boston University Student Responsibilities: `https://www.bu.edu/dos/policies/student-responsibilities/`
- All Spark! project teams
  - Project Managers: These are the project leads and will communicate with the client directly, they will assist with administrative support (meeting scheduling, agenda setting), and will be a point of contact for project questions / concerns. Project Managers are also responsible for grading all Spark! project deliverables as detailed in the syllabus below.
  - Team Lead: These students will assist the Project Manager in attending client meetings, organizing team questions, and facilitating team meetings.
  - Team Members: These students work collaboratively with each other on the project goals.

For details on what you must submit as part of your project, see the section "Project Deliverables" at the end of this syllabus.

## Spark! Collaboration

BU Spark! offers students an opportunity to work on technical projects provided by companies or organizations in the Greater Boston area through our experiential learning lab (X-Lab). For this semester, Spark! has partnered with DS701 to offer a diverse selection of external data science projects scoped to support the course's learning outcomes and enhance the student experience. To learn more about Spark!, please visit their website: `https://www.bu.edu/spark/`.

Your project team will be led by one of the Spark! Project managers. Their role is to support the student team's work plan, manage client communication and expectations, organize weekly and biweekly meetings, and to oversee project deliverable grading.

Spark! projects are a great opportunity for students to get real-world project experience to highlight on their github and CV. These projects have already been curated and will be presented during "Pitch Day". Project descriptions will be made available at the start of the semester.

## Academic Honesty

You may discuss homework assignments with classmates, but you are solely responsible for what you turn in. Collaboration in the form of discussion is allowed, but all forms of cheating (copying parts of a classmate's assignment, plagiarism from books or old posted solutions) are NOT allowed. We – both teaching staff and students – are expected to abide by the guidelines and rules of the Academic Code of Conduct (which is at `http://www.bu.edu/academics/policies/academic-conduct-code/`).

Graduate students must also be aware of and abide by the GRS Academic Conduct code at `http://www.bu.edu/cas/students/graduate/forms-policies-procedures/academic-discipline-procedures/`.

You can probably, if you try hard enough, find solutions for homework problems online. Given the nature of the Internet, this is inevitable. Let me make a couple of comments about that:

1. If you are looking online for an answer because you don't know how to start thinking about a problem, talk to Ms. Lu or myself, who may be able to give you pointers to get you started. Piazza is great for this – you can usually get an answer in an hour if not a few minutes.

2. If you are looking online for an answer because you want to see if your solution is correct, ask yourself if there is some way to verify the solution yourself. Usually, there is. You will understand what you have done *much* better if you do that. So ... it would be better to simply submit what you have at the deadline (without going online to cheat) and plan to allocate more time for homeworks in the future.

## Course Schedule

| Date | Topics | Assignments Due |
|---|---|---|
| 9/5 | MS Student Orientation | |
| 9/7 | Introduction and Essential Tools | |
| 9/12 | Spark Pitch Day | |
| 9/14 | Distance and Similarity Functions, Timeseries | Spark Project Choices |
| 9/19 | Clustering I: k-means | |
| 9/21 | Clustering II: In practice | Homework 0 |
| 9/26 | Clustering III: Hierarchical Clustering | |
| 9/28 | Clustering IV: GMM and Expectation Maximization | |
| 10/3 | Classification I: Decision Trees | |
| 10/5 | Classification II: *k*-Nearest Neighbors | Homework 1 |
| 10/10 | Classification III: Naive Bayes, HMMs | |
| 10/12 | Support Vector Machines | Project Deliverable 0 |
| 10/15 | | MIDTERM START |
| 10/17 | Ethical Analysis of Data Science Projects (Seth Villegas) | |
| 10/19 | SVD I : Low Rank Approximation | |
| 10/24 | SVD II: Dimensionality Reduction | |
| 10/26 | Nonlinear Dimensionality Reduction | MIDTERM END midnight |
| 10/31 | Regression I: Linear Regression | Project Deliverable 1 |
| 10/31 | | Project Ethics Audit |
| 11/2 | Regression II: Logistic Regression | |
| 11/7 | Regression III: In Practice | Project Deliverable 2 |
| | | Homework 2 |
| 11/9 | Recommender Systems | |
| 11/14 | **Early Insight Presentations** | |
| 11/16 | Gradient Descent | Project Deliverable 3 |
| 11/21 | Neural Networks I | Homework 3 |
| 11/23 | NO CLASS; Thanksgiving Break | |
| 11/28 | Neural Networks II | Project Deliverable 4 |
| 11/30 | Network Analysis I | |
| 12/5 | Network Analysis II | |
| 12/7 | **Final Project Presentations** | |
| 12/12 | **Final Project Presentations** | Final Project Report |
| 12/13 | **Demo Day** | Participation Required |

# Project Deliverables

See: `https://github.com/BU-Spark/`. Project deliverables can be modified with approval of PMs, the client, or the instructor.

## Project Deliverable 0

Teams should have set up weekly meetings with their client for the remainder of the semester, reviewed the project scope, and submitted a pull request with the revised and final project description. Project descriptions should include data sources that your team will collect, including any additional datasets you identify that you think would enhance the project, specific questions that will be answered and the step by step approach you will take for transforming the data (cleaning) and answering strategic questions.

**Checklist**

1. Reviewed all previous material
2. Revised scope of the project if needed
3. Identify / list limitations with data and potential risks of achieving project goal
4. Meet with client to review the project
5. Schedule weekly meetings with PMs and bi-weekly with client
6. Submit a PR with the revised project proposal including list of limitations

## Project Deliverable 1

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request. If data has already been collected for your project you must answer two questions.

**Checklist**

1. Collect and pre-process a preliminary batch of data
2. Perform a preliminary analysis of the data
3. Answer one key question
4. Refine project scope and list of limitations with data and potential risks of achieving project goal
5. Submit a PR with the above report and modifications to original proposal

## Project Deliverable 2

More data should have been collected to perform a more thorough analysis of the data and attempt to answer one additional question relevant to your project proposal which you will submit as a pull request.

**Checklist**

1. Collect and pre-process a secondary batch of data
2. Refine the preliminary analysis of the data performed in PD1
3. Answer another key question
4. Refine project scope and list of limitations with data and potential risks of achieving project goal
5. Submit a PR with the above report and modifications to original proposal

## Project Deliverable 3 (v1 Final Report)

All data should have been collected. All project questions should have been reviewed, answered, and submitted in a written document outlining findings as a PR. You will also be asked to submit the associated data and a README explaining what each label/feature in your dataset represents. Your team should meet with the client before this deliverable.

**Checklist**

1. All data is collected
2. Refine the preliminary analysis of the data performed in PD1 & 2
3. Answer another key question
4. Attempt to answer overarching project question
5. Create a draft of your final report
6. Refine project scope and list of limitations with data and potential risks of achieving project goal
7. Submit a PR with the above report and modifications to original proposal

## Project Deliverable 4 (v2 Final Report)

This is a draft of your final report that has been reviewed by your client. It includes all visualizations, results, data, and code up to this point, along with proper documentation on how to reproduce your results, compile and use your codebase, and navigate your dataset. Your team will submit this as a PR.

## Final Project Deliverable

This should be an enhancement of deliverable 4.