DS-320: Algorithms for Data Science

Boston University

Fall 2024

Teaching Staff

Instructors:

- Prof. Krzysztof Onak (konak@bu.edu)
- Prof. Aldo Pacchiano (pacchian@bu.edu)

TFs/TAs:

- Anup De (anupde@bu.edu)
- Xinyi "Selena" Hu (xhu07@bu.edu)
- Kevin Quinn (quinnk@bu.edu)

Course Assistants:

- Jin Kyu "Jaden" Cho
- Johnathan Finizio
- Jason Huang
- Chenqi Wang

The preferred way to communicate with the teaching staff is via Piazza. Use email only for topics you are not comfortable discussing via Piazza or if you do not receive a reply for an extended period of time.

Lectures and Discussion Sections

You are expected to attend the lecture and discussion section for which you registered.

Section A:

- **Lecture (A1):** Prof. Onak, Tuesday/Thursday 2–3:15pm, CDS B64
- **Discussion section (A2):** Wednesday 10:10–11am, CGS 315
- **Discussion section (A3):** Wednesday 11:15am–12:05pm, CGS 315

Section B

- **Lecture (B1):** Prof. Pacchiano, Tuesday/Thursday 9–10:15am, CGS 527
- **Discussion section (B2):** Wednesday 2:30–3:20pm, PSY B37
- **Discussion section (B3):** Wednesday 3:35–4:25pm, PSY B37

Office Hours

You are invited to attend any office hours, independently of the lecture and discussion section for which you registered. The current schedule of office hours is

- Prof. Krzysztof Onak: CDS 1443, Tuesdays, 10:45am–12:45pm
- Prof. Aldo Pacchiano: CDS 1411, Thursdays, 11am–1pm
- CDS 13th floor, blue northwest corner, Mondays, 1–3pm
- CDS 5th floor pavilion, Wednesdays, 5–7pm
- CDS 13th floor, blue northwest corner, Thursdays, 1–3pm
- CDS 13th floor, blue northwest corner, Fridays, 1–3pm

For more up to date information, see the pinned "Office Hours" post (@6) on Piazza.

Course Content and Objectives

Official course description: This course covers the fundamental principles underlying the design and analysis of algorithms. We will walk through classical design methods, such as greedy algorithms, design and conquer, and dynamic programming, focusing on applications in data science. We will also study algorithmic methods more specific to data science and machine learning. The course places a particular emphasis on algorithmic efficiency, crucial with large and/or streaming data sets, for which multiple scans of data are infeasible, including the use of approximation and randomized algorithms.

Learning objectives: In this course, we want you to

- develop creative problem solving and algorithmic thinking skills,
- learn a repertoire of universal algorithmic techniques,
- get comfortable understanding and writing formal definitions and statements,
- write clear and convincing arguments,
- learn how to identify and formally describe an algorithmic problem underlying a given applied problem,
- develop skills useful for real–world job interviews (which often involve questions such as "Can you solve this algorithmic problem in Python?").

BU Hub: Quantitative Reasoning II and Critical Thinking[1]

Prerequisites

Please consult the official data science program prerequisites—there may have been some changes—for an up–to–date version if you are a data science major or minor.

[1] If fulfilling these requirements is important for you, please crosscheck that nothing has changed in this regard.

In terms of actual technical competencies you should have, keep in mind that this is a theoretical problem–solving and proof–writing course. Whatever background you are coming from, your preparation should include:

- **Programming:** We won't directly program in this class, but understanding what programming is and the ability to read and write pseudocode[2] is crucial. Future editions of the course may require solving some homework questions in a real programming language. In the past, some students of DS-320 already found it useful to implement their solutions to experimentally verify that they are correct.
- **Data structures:** This course mostly focuses on algorithms and assumes knowledge of basic data structures (queues, stacks, priority queues such as binary heaps, binary search trees, etc.). We will review some of the useful material or briefly mention it to the extent it is needed for the topics we discuss. Everything you need to know in advance should have been covered in DS-210 if you are a data science major.
- **Math:** Basic calculus, combinatorics, probability, and linear algebra are all very useful. A basic understanding what constitutes a proof and the ability to write simple proofs is crucial. This is something that we will practice and improve on in this class. Any more advanced class that gives you training in this could be very useful.

Course information and tools

- Piazza (materials, announcements, and discussions):
  `https://piazza.com/bu/fall2024/ds320/home`
  Access code: `023sd` (to be disabled after the registration period ends)
- Gradescope entry code (homework submissions): `Z3JV5P`

Course Requirements

The final grade will be a combination of the following factors:

- class participation (in particular, asking and answering questions in class and on Piazza): 5%
- weekly problem sets: 38% overall
- two midterms: 19% each
- final: 19%

[2]Pseudocode is an informal language in which we will describe our algorithms. It roughly resembles any modern structured programming language with subroutines, loops (for/while/...), conditional statements, etc. Whenever you are asked to write a solution in psedocode and you are not confident how to express your ideas in our informal pseudocode, you are allowed to use any mainstream programming language such as Python instead. Note, however, that the whole point of pseudocode is to make things simpler, so writing in pseudocode should be simpler than using an actual programming language.

Course participation. For best learning outcomes, it is very important to actively participate in both lectures and discussion sections. To encourage this, we will keep track of virtual participation cards. When you participate in good faith during class, we will collect your card. Every now and then (probably roughly every two weeks), we will announce that the cards have been distributed back. We will keep track of the number of times the participation card was collected for each student on Gradescope. Your credit for class participation will be a (possibly non-linear) function of the number of times your card was collected and may include other factors as well.

Weekly problem sets.

- Expect to spend 10 hours per week on homework.
- **Late policy:** Homeworks will typically be due on Wednesdays at 11:59pm. You can submit your solution up to one day late (typically 11:59pm on Thursday) but you may lose 10% of credit for this.
- Your lowest homework score will be dropped at the end of the semester.
- Your written assignments must be prepared with LaTeX (or other software with the same functionality), not handwritten. You may use Overleaf.
- You must submit your homework via Gradescope in the PDF format. Please select the pages in your solution that correspond to each problem.
- Regrade requests must be submitted within 14 days of receiving the graded assignment and only via Gradescope. You must submit an explanation which problems were graded incorrectly and an argument why the solution is in fact correct. If you submit a regrade request, you accept that the entire assignment or exam may be regraded, not just the problem that you believe was graded incorrectly.

Midterms and the final exam. The course will have two midterms and a final exam:

- **Midterm exam 1:** evening exam, October 8, 6–8pm, PHO 206
- **Midterm exam 2:** evening exam, November 7, 6–8pm, CGS 130
- **Final exam:** to be scheduled

Please make sure you do not make any travel arrangements before the specific dates above are confirmed. In particular, please wait for us to confirm the date of the final exam, even after it appears in the BU scheduling system (because we may opt out for a single combined exam for both sections). All exams will be closed book.

Homework collaboration policy. You are allowed to collaborate on your homework with up to three of your classmates. However the assignments you hand in should be written up by yourself and represent your own work and thoughts. In particular, you are allowed to discuss ideas with them in person, but as a rough rule, nobody should leave the room with anything written down. If you really understand the discussion, you should be able to reconstruct it on your own.

Your must list your collaborator's names on the top of your assignment. If you don't work with anyone, you must write "Collaborators: none." Not listing your collaborators may result in a credit reduction.

Academic code of conduct. You have to adhere to BU's academic conduct policy:
https://www.bu.edu/academics/policies/academic-conduct-code/

Generative AI and grammar tools. Using generative AI tools such as ChatGPT or Google Bard is, generally, not allowed for this class. See also the CDS GAIA policy:

https://www.bu.edu/cds-faculty/culture-community/gaia-policy/

It is allowed, however, to use modern grammar and language verification tools that go beyond the traditional spell checking.

Grade cutoffs

We will determine grade cutoffs after all assignments and exams have been graded. Grade cutoffs will take into account my assessment of the difficulty level of the assignments and exams, and my assessment of what is expected for each letter grade.

Lecture recording

In this course, your voice and image may be recorded by the instructor to share video recordings of lectures with the class participants for instructional purposes. (Note that it has not been decided yet whether video recordings will be made this year for either of the sections. This depends on logistics of a specific classroom and other considerations.) These recordings are not to be shared outside of the classroom without permission of persons being recorded. The recordings may be retained later to improve the quality of future educational offerings but they will not be posted publicly.

Additionally, some students may be allowed to record lectures as a disability accommodation. Sharing these recordings without permission of class participants is not allowed and they should be deleted at the end of the semester. If you have this kind of accommodation, you should notify the lecturer and receive the permission to record audio.

If this lecture recording policy makes you uncomfortable, please discuss it with the instructor.

Materials

There is no required textbook. There will be suggested readings from various textbooks and lecture notes during the course. If you prefer to use an algorithmic textbook, two popular books are "Algorithm Design" by Kleinberg and Tardos and "Introduction to Algorithms" by Cormen, Leiserson, Rivest, and Stein. We will also share links to other materials related to the course, and you are encouraged to share links to general resources you found valuable on Piazza.

Reasonable accommodations

If you are a student with a disability or believe you might have a disability that requires accommodation, please contact the Office for Disability Services at 617-353-3658 or access@bu.edu. Please also notify the instructor about any accommodation that you may require as soon as possible. We may not be able to provide some accommodations if we do not learn about them sufficiently early.

Tentative topics to be covered

This is a highly tentative list of topics that will be covered. Anything beyond the basics may be replaced by other topics. We mark some topics below as "(optional)" to indicate that they are more likely to not be covered. Additionally, the parallel sections of the class may cover slightly different sets of topics.

- Overview and policies. Runtime and asymptotics.
- Insertion Sort, induction, and the comparison–based lower bound.
- Abstract data types and Depth–First Search (DFS).
- More DFS, including Topological Sort.
- Breadth–First Search (BFS) and testing bipartiteness.
- Greedy algorithms I: Dijkstra's algorithm.
- Greedy algorithms II: optimal caching.
- Greedy algorithms III: scheduling.
- Greedy algorithms IV: Huffman codes.
- Divide & Conquer I: Mergesort and solving recurrences.
- Divide & Conquer II: closest pair of points.
- Divide & Conquer III: integer and matrix multiplication.
- (optional) Faster graph algorithms via matrix multiplication.
- Dynamic Programming I: weighted interval scheduling.
- Dynamic Programming II: segmented least squares.
- Dynamic Programming III: knapsack.
- (optional) Dynamic Programming IV: optimal order of matrix multiplications
- (optional) Dynamic Programming V: Bellman–Ford algorithm.
- NP–completeness.
- Linear Programs I: introduction.
- (optional) Linear Programs II: examples and duality.
- (optional) Linear Programs III: more duality.
- (optional) Zero sum games and the minimax theorem.
- (optional) Multiplicative weight update.
- (optional) Stable matching.
- (optional) Approximation algorithms: random and online.
- (optional) Sample streaming and big data algorithms.

Tentative schedule of lectures, exams, and homework

September 3: Lecture 1
September 5: Lecture 2

September 10: Lecture 3
September 11: Homework 1 out
September 12: Lecture 4

September 17: Lecture 5
September 18: Homework 1 due, Homework 2 out
September 19: Lecture 6

September 24: Lecture 7
September 25: Homework 2 due, Homework 3 out
September 26: Lecture 8

October 1: Lecture 9
October 2: Homework 3 due
October 3: Lecture 10

October 8: no lecture, evening exam 6–8pm in PHO 206
October 8: Homework 4 out
October 10: Lecture 11

October 15: no lecture, substitute Monday at BU
October 16: Homework 4 due, Homework 5 out
October 17: Lecture 12

October 22: Lecture 13
October 23: Homework 5 due, Homework 6 out
October 24: Lecture 14

October 29: Lecture 15
October 30: Homework 6 due
October 31: Lecture 16

November 5: Lecture 17
November 6: Homework 7 out
November 7: no lecture, evening exam 6–8pm in CGS 130

November 12: Lecture 18
November 13: Homework 7 due, Homework 8 out
November 14: Lecture 19

November 19: Lecture 20
November 20: Homework 8 due, Homework 9 out
November 21: Lecture 21

November 26: Lecture 22
November 27: no discussion sections (Thanksgiving recess)
November 28: no lecture (Thanksgiving recess)

December 3: Lecture 23
December 4: Homework 9 due
December 5: Lecture 24

December 10: Lecture 25
December 11–13: study period

December 16–20: final exams, DS-320 exam(s) to be scheduled