COMPSCI 514: Algorithms for Data Science Syllabus

All information on this syllabus, along with a detailed outline of topics covered and a tentative schedule can be found on the course website https://people.cs.umass.edu/~mcgregor/CS514S22 .

Logistics

Time and Location: Tu/Th 11:30am–12:45pm in Hasbrouck 126.

Attendance Policy: All slides will be available at the above website. Attendance at lectures is recommended but if you have to miss lectures for whatever reason, that's okay and will to affect your grade.

Staff:

- Instructor: Andrew McGregor (mcgregor@cs.umass.edu)
- TAs: Shiv Shankar (sshankar@cs.umass.edu) and Thuy Trang (Weronika) Nguyen (thuytrangngu@umass.edu)
- See Moodle for links and the times of office hours.

Course Content

Course Description: With the advent of social networks, ubiquitous sensors, and large-scale computational science, data scientists must deal with data that is massive in size, arrives at blinding speeds, and often must be processed within interactive or quasi-interactive time frames. This course studies the mathematical foundations of big data processing, developing algorithms and learning how to analyze them. We explore methods for sampling, sketching, and distributed processing of large scale databases, graphs, and data streams for purposes of scalable statistical description, querying, pattern mining, and learning. Course was previously COMPSCI 590D. 3 credits.

Prerequisites: The most important undergraduate prerequisites are COMPSCI 240 (Probability) and COMPSCI 311 (Algorithms). This is a theoretical course with an emphasis on algorithm design, correctness proofs, and analysis. Aside from a general background in algorithms, a strong mathematical background, particularly in linear algebra and probability is required.

Textbooks: This is no official textbook for this class. We will use some material two books, available as free pdfs linked to below. I will post readings from these books and other sources on the course website before class when relevant.

- Foundations of Data Science, Avrim Blum, John Hopcroft and Ravi Kannan.
- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman and Jeff Ullman.

You may also find some helpful reference material in these similar classes taught at other universities.

- The Modern Algorithmic Toolbox, Gregory Valiant at Stanford.
- Sketching Algorithms for Big Data, Piotr Indyk and Jelani Nelson at MIT/Harvard.
- Algorithmic Techniques for Big Data, Moses Charikar at Stanford.

Approximate Schedule:

1. Course overview. Probability review. (1 lecture)
2. Randomized Methods and Sketching (Roughly 12 lectures):
   - Estimating set size by counting duplicates. Concentration Bounds: Markov's inequality. Random hashing for efficient lookup.
   - 2-universal and pairwise independent hashing. Hashing for load balancing.
   - Concentration Bounds Continued: Chebyshev's inequality. The union bound. Exponential tail bounds (Bernstein's inequality).
   - Central limit theorem. Bloom filters and their applications.
   - Streaming algorithms: Min-Hashing for distinct elements.
   - Flajolet-Martin and HyperLogLog. Jaccard similarity estimation with MinHash for audio fingerprinting, document comparison, etc. Start on locality sensitive hashing and nearest neighbor search.
   - Similarity search. SimHash for Cosine similarity.
   - The frequent elements problem and count-min sketch.
   - Dimensionality reduction, low-distortion embeddings, and the JL Lemma.
   - Example application to clustering. Connections to high-dimensional geometry.
   - Midterm Review
3. Spectral Methods (Roughly 7 lectures):
   - Intro to principal component analysis, low-rank approximation, data-dependent dimensionality reduction.
   - Projection matrices and best fit subspaces.
   - Optimal low-rank approximation via eigendecomposition. Principal component analysis.
   - SVD and applications of low-rank approximation beyond compression. Matrix completion, LSA, and word embeddings.
   - Linear algebraic view of graphs. Spectral graph partitioning and clustering.
   - Stochastic block model.
   - Computing the SVD: power method, Krylov methods. Connection to random walks and Markov chains.
4. Optimization (Roughly 4 lectures):
   - Optimization and gradient descent analysis for convex functions.
   - Constrained optimization and projected gradient descent.
   - Online learning and regret. Online gradient descent.
   - Finish up online gradient descent and stochastic gradient descent analysis. Course conclusion/review.

Assessment and Participation

Piazza: We will use Piazza for class discussion, questions, and general class communications.

Grading:

- Problem Sets: 30%, weighted equally.
- Weekly Quizzes: 15%, weighted equally.
- Midterm: 25%.
- Final: 25%.
- Participation: 5%. This is based on asking good questions and answering the questions of others on the Piazza Forum.

Grade Cut Offs: A: 90%, **A-:** 88%, **B+:** 86%, **B:** 80%, **B-:** 78%, **C+:** 76%, **C:** 70%.

Problem Sets: Problem sets can be completed in **groups of up to three students**. If you work in a group, you submit a single problem set together. You may not talk to people outside your group (other than the instructor and TA) about the homework until solutions are posted. We very strongly encourage you to work in a three person group, as it will give an advantage in doing the problem sets. At the beginning of the semester we will make a Piazza post where you can look for teammates. We will also have random breakouts during lecture so that you can get to know some of your classmates. Problem set submissions will be via Gradescope. If working in a group, only one member of each group should submit the problem set, marking the other members in the group as part of the submission in Gradescope.

- The entry code for Gradescope will be posted on Moodle.
- No late homework submissions will be accepted.
- I strongly encourage students to type up problem sets using LaTeX. An optional LaTeX template for problem sets can be downloaded here. While they may seem cumbersome at first, these tools will save you a lot of time in the long run!

Weekly Quizzes: A quiz will be posted on Piazza each Friday afternoon, due the following Monday at 8pm. These are open note quizzes (designed to take less than an hour) to check that you are following the material and help me make adjustments if needed. Quizzes will include check-in questions asking for feedback on class pacing and on topics that need clarification, or that you would like to see discussed more. Quizzes should be completed on your own, without collaboration. No late quizzes will be accepted but we'll drop your lowest quiz (so if you miss one quiz, it's no big deal.)

Exams: We will have a midterm in March (provisionally 10th or 11th March) and final in May (see `https://www.umass.edu/registrar/students/courses-and-schedules/final-exam-info`). Each exam will last 2 hours.

Academic Honesty

For any cheating on problem sets, quizzes, or other assignments, students will receive a 0% on the assignment for the first violation, and fail the class for a second violation. *Any cheating on the midterm or final will lead to failing the class.* For fairness, we apply these rules universally, without exceptions. Cheating includes but is not limited to: close collaboration with students outside your problem set group on problem sets, searching for solutions online, any collaboration on quizzes/exams, and any consultation with solution sets or problem sets obtained from students that have taken the course in the past.

Disability Services and Accommodations

UMass Amherst is committed to making reasonable, effective, and appropriate accommodations to meet the needs to students with disabilities and help create a barrier-free campus. If you have a documented disability on file with Disability Services, you may be eligible for reasonable accommodations in this course. If your disability requires an accommodation, please notify me within the first two weeks of the course so that we may make arrangements in a timely manner. I understand that people have different learning needs, home situations, etc. If something isn't working for you in the class, please reach out and we'll try to work it out.

Helpful UMass Resources

- Academic Honesty Policy
- English as a Second Language (ESL) Program
- Center for Counseling and Psychological Health

Learning Objectives

1. Students will learn about modern tools for data processing, including random sampling and hashing, low-memory streaming algorithms, linear and non-linear dimensionality reduction, spectral graph theory, and continuous optimization. A major goal is to be familiar at a high level with a breadth of algorithmic tools beyond combinatorial algorithms, which are the main focus of most undergraduate algorithms courses.
2. Through problem sets, students will develop the ability to apply and modify these algorithmic tools to tackle new problems, beyond those discussed in class. They will strengthen their ability to think creatively about algorithmic problems and push beyond known approaches, to develop solutions of their own.
3. Through assessments that emphasize formal proofs, students will strengthen their ability to formulate problems mathematically and analyze them rigorously.
4. Through algorithmic problems, students will practice applying fundamental tools in probability theory and linear algebra, which are broadly applicable in data science and machine learning. These include concentration bounds and methods for decomposing complex random variables, eigendecomposition, orthogonal projection, important matrix identities, and fundamentals of high-dimensional geometry and random matrix theory.