COMPSCI 514: Algorithms for Data Science

Prof. Cameron Musco

University of Massachusetts Amherst. Fall 2022.

Lecture 1

Motivation For this Class

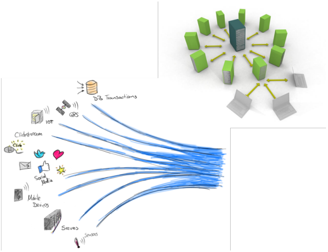
The ability to analyze and learn from massive datasets is critical across many industries, the sciences, and beyond.

- Twitter receives 6,000 tweets per second, 500 million/day. Google receives 60,000 searches per second, 5.6 billion/day.
  - How do they process them to target advertisements? To predict trends? To improve their products?
- The Large Synoptic Survey Telescope will take high definition photographs of the sky, producing 15 terabytes of data/night.
  - How do they denoise and compress the images? How do they detect anomalies such as changing brightness or position of objects to alert researchers?

A New Paradigm for Algorithm Design

- Traditionally, algorithm design focuses on fast computation when data is stored in an efficiently accessible centralized manner (e.g., RAM on one machine).
- Massive data sets require storage in a distributed manner or processing in a continuous stream.

vs.



- Even 'simple' problems can become very difficult in this setting.

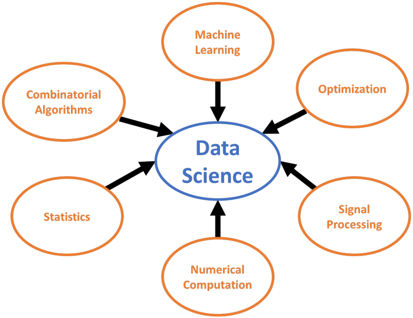
A New Paradigm for Algorithm Design

For example:

- How can Twitter rapidly detect if an incoming Tweet is an exact duplicate of another Tweet made in the last year? Given that no machine can store all Tweets made in a year.
- How can Google estimate the number of unique search queries that are made in a given week? Given that no machine can store the full list of queries.
- When you use Shazam to identify a song from a recording, how does it provide an answer in < 10 seconds, without scanning over all $\sim$ 8 million audio files in its database.

Motivation for This Class

A Second Motivation: Data Science is highly interdisciplinary.



- Many techniques that aren't covered in the traditional CS algorithms curriculum.
- Emphasis on building comfort with mathematical tools that underly data science and machine learning.

What We'll Cover

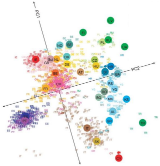
Section 1: Randomized Methods & Sketching

How can we efficiently compress large data sets in a way that lets us answer important algorithmic questions rapidly?

- Probability tools and concentration inequalities.
- Randomized hashing for efficient lookup, load balancing, and estimation. Bloom filters.
- Locality sensitive hashing and nearest neighbor search.
- Streaming algorithms: identifying frequent items in a data stream, counting distinct items, etc.
- Random compression of high-dimensional vectors: the Johnson-Lindenstrauss lemma, applications, and connections to the weirdness of high-dimensional geometry.

What We'll Cover

Section 2: Spectral Methods

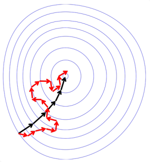


How do we identify the most important features of a dataset using linear algebraic techniques?

- Principal component analysis, low-rank approximation, dimensionality reduction.
- The singular value decomposition (SVD) and its applications to PCA, low-rank approximation, LSI, MDS, …
- Spectral graph theory. Spectral clustering, community detection, network visualization.
- Computing the SVD on large matrices via iterative methods.

If you open up the codes that are underneath [most data science

Section 3: Optimization



Fundamental continuous optimization approaches that drive methods in machine learning and statistics.

- Gradient descent. Analysis for convex functions.
- Stochastic and online gradient descent.
- Focus on convergence analysis.

A small taste of what you can find in COMPSCI 590OP or 690OP.

Important Topics We Won't Cover

- Systems/Software Tools.
  - COMPSCI 532: Systems for Data Science
- **Machine Learning/Data Analysis Methods and Models.**
  - E.g., regression methods, kernel methods, random forests, SVM, deep neural networks.
  - COMPSCI 589/689: Machine Learning

Style of the Course

This is a theory course.

- Build general mathematical tools and algorithmic strategies that can be applied to a wide range of problems.
- Assignments emphasize algorithm design, correctness proofs, and asymptotic analysis (relatively little required coding).
- The homework will push beyond what is taught in class. You will get stuck, and not see the solutions right away. This is a key way to build mathematical and algorithm design skills.
- A strong algorithms and mathematical background (particularly in linear algebra and probability) **are required**.
- Prereqs: COMPSCI 240 and COMPSCI 311. If you are an MS student and unsure about your background, email me or come chat.

For example: Baye's rule in conditional probability. What it means for a vector x to be an eigenvector of a matrix A, orthogonal projection, greedy algorithms, divide-and-conquer algorithms.

Course Logistics

See course webpage for logistics, policies, lecture notes, assignments, etc.:

http://people.cs.umass.edu/~cmusco/CS514F22/

See Moodle page for this link if you lose it, or search my name and follow the link from my homepage.

Moodle will be used for weekly quizzes, but the course page for mostly everything else.

Personnel

Professor: Cameron Musco

- Email: cmusco@cs.umass.edu
- Office Hours: Over Zoom, Tuesdays, 2:30pm-3:30pm (directly after class) in CS 234.
- I encourage you to come as regularly as possible to ask questions/work together on practice problems.
- If you need to chat individually, please email meet to set up a time.

TAs:

- Forsad Al Houssain
- An La
- Mohit Yadav

See website for office hours and contact info.

Online Section

There is also an online version of 514 taught this semester by Andrew McGregor, Tue/Thu 11:30am-12:45pm.

- The sections will closely parallel each other, and share the same TAs.
- You may attend Prof. McGregor's lectures on Zoom if it is helpful.
- See Moodle for the Zoom link.

Piazza and Participation

We will use Piazza for class discussion and questions.

- See website for link to sign up.

You may earn up to 5% extra credit for participation.

- Asking good clarifying questions and answering questions during the lecture or on Piazza.
- Actively participating in office hours.
- Answering other students' or instructor questions on Piazza.
- Posting helpful links on Piazza, e.g., resources that cover class material, research articles related to the class, etc.
- It is completely fine to post private questions on Piazza, but these don't count towards participation credit.
- You can post anonymously on Piazza. Instructors will see the author behind all posts, so we can assign participation credit.

Textbooks and Materials

We will use material from two textbooks (links to free online versions on the course webpage): *Foundations of Data Science* and *Mining of Massive Datasets*, but will follow neither closely.

- I will post optional readings a few days prior to each class.
- Lecture notes will be posted before each class, and annotated notes posted after class.
- Recordings of the live lectures will also be posted on Echo360.
- Sometimes it takes a lecture or two to get the Echo360 set up working properly.

Homework

We will have 5 problem sets, which you may complete in **groups of up to 3 students**.

- We strongly encourage working in groups, as it will make completing the problem sets much easier/more educational.
- Collaboration with students outside your group is limited to discussion at a high level. You may not work through problems in detail or write up solutions together.
- See Piazza for a thread to help you organize groups.

Problem set submissions will be via Gradescope.

- See website for a link to join. **Entry Code: 2KBPNG**

Weekly Quizzes

We will release an online quiz in Moodle each Thursday after lecture, due the next Monday at 8pm.

- Designed as a check-in that you are following the material, and to help me make adjustments as needed.
- Will take around 15-30 minutes per week, open notes.
- Will also include free response check-in questions to get your feedback on how the course is going, what material from the past week you find most confusing, interesting, etc.

Grading

Grade Breakdown:

- Problem Sets (5 total): 40%, weighted equally.
- Weekly Quizzes: 10%, weighted equally.
- Midterm (October 20th, in class): 25%.
- Final (December 14th, 10:30am - 12:30pm): 25%.
- Extra Credit: Up to 5% for participation, and more available on problem sets and exams.

Academic Honesty:

- A first violation cheating on a homework, quiz, or other assignment will result in a 0 on that assignment.
- A second violation, or cheating on an exam will result in failing the class.
- For fairness, I adhere very strictly to these policies.

Disability Services and Accomodations

UMass Amherst is committed to making reasonable, effective, and appropriate accommodations to meet the needs to students with disabilities.

- 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 email me by **next Thursday 9/15** so that we can make arrangements.

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 let's try to work it out.

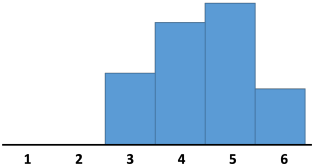
Questions?

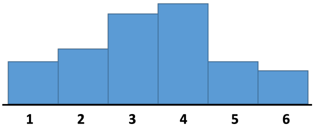
Section 1: Randomized Methods & Sketching

Some Probability Review

Consider a random X variable taking values in some finite set $S \subset \mathbb{R}$. E.g., for a random dice roll, $S = \{1, 2, 3, 4, 5, 6\}$.

- **Expectation:** $\mathbb{E}[X] = \sum_{s \in S} \Pr(X = s) \cdot s$.
- **Variance:** $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.





Exercise: Show that for any scalar α, $\mathbb{E}[\alpha \cdot X] = \alpha \cdot \mathbb{E}[X]$ and $\text{Var}[\alpha \cdot X] = \alpha^2 \cdot \text{Var}[X]$.

Independence

Consider two random events A and B.

- **Conditional Probability:**

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

- **Independence:** A and B are independent if:

$$\Pr(A|B) = \Pr(A).$$

Using the definition of conditional probability, independence means:

$$\frac{\Pr(A \cap B)}{\Pr(B)} = \Pr(A) \implies \Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

$A \cap B$: event that both events A *and* B happen.

For Example: What is the probability that for two independent dice rolls the first is a 6 and the second is odd?

Independent Random Variables: Two random variables X, Y are independent if for all s, t, $X = s$ and $Y = t$ are independent events. In other words:

$$\Pr(X = s \cap Y = t) = \Pr(X = s) \cdot \Pr(Y = t).$$

Linearity of Expectation and Variance

Think-Pair-Share: When are the expectation and variance linear?

I.e., under what conditions on X and Y do we have:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

and

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y].$$

X, Y: any two random variables.

Linearity of Expectation

$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for any random variables X and Y.

Proof:

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_{s \in S} \sum_{t \in T} \Pr(X = s \cap Y = t) \cdot (s + t) \\
&= \sum_{s \in S} \sum_{t \in T} \Pr(X = s \cap Y = t) \cdot s + \sum_{t \in T} \sum_{s \in S} \Pr(X = s \cap Y = t) \cdot t \\
&= \sum_{s \in S} \Pr(X = s) \cdot s + \sum_{t \in T} \Pr(Y = t) \cdot t
\end{aligned}$$

(law of total probability)

$$= \mathbb{E}[X] + \mathbb{E}[Y].$$