CS-119 Syllabus

Contact Info

Professor	Jitendra Singh	617-444-9640
TA	Hiba Eltigani	

Meetings

Lectures: Mon, Wed 10:30 am — 11:45 am from Jan 17, 2024 to May 1, 2024[1].
Joyce Cummings Center, Room 180 .

Course Description

Big Data deals with emerging applications in science and engineering disciplines that generate and collect data at unprecedented speed, scale, and complexity that need to be managed and analyzed efficiently.

This course introduces the latest techniques and infrastructures developed for big data including parallel and distributed database systems, map-reduce infrastructures, scalable platforms for complex data types, stream processing systems, and cloud-based computing.

The course content will be a blend of theory, algorithms and practical (hands on) work, involving software design, coding, testing and debugging!

Prerequisites: Formally, there are no prerequisites for this course. Still, CS-119 should not be your first programming course. Familiarity with database internals (CS-115) is helpful but is not required. The Background Assessment exercise available in Gradescope will give you an indication of your preparedness for CS-119.

[1] The course will also meet on Thursday, February 22, per the University calendar.

Programming Languages: Big Data work across the industry involves Python, Java, Scala, shell programming and SQL. Most of our work in this class will be in Python, but familiarity with the programming languages cited will give you a head start. It is expected that students taking CS-119 know at least some of these and will pick up the rest, as required, on their own.

Theme: Some of the exercises in this course playfully evoke *Alice in Wonderland* by Lewis Carroll[2].

Textbooks on Big Data Theory

The Datacenter as a Computer *Designing Warehouse-Scale Machines*, Third Edition Luiz André Barroso Urs Hölzle Parthasarathy Ranganathan	Springer Cham Copyright: 2019 ISBN: 978-3031006333 Downloadable from Tufts Library with your credentials.
Mining of Massive Datasets, 2nd edition Jure Leskovec Anand Rajaraman Jeff Ullman	Cambridge University Press Copyright: 2014 ISBN: 978-1107077232 Download for private use only from Stanford InfoLab.
Everybody lies *Big data, new data, and what the internet can tell us about who we really are* Seth Stephens-Davidowitz	Harper Collins Copyright: 2017 ISBN: 9780062390875
Data Science in Context *Foundations, Challenges, Opportunities* Alfred Z. Spector Peter Norvig Chris Wiggins Jeannette M Wing	Cambridge University Press Copyright: October 20, 2022 ISBN: 978-1009272209 Download manuscript for private use only from the publisher

[2] In addition to authoring Alice in Wonderland, Lewis Carroll was a Mathematician. He is credited with a paper on infinite logic, What the Tortoise said to Achilles, popularized by Douglas Hofstadter's writings.

Hofstadter's *I am a Strange Loop* and its predecessor *Gödel, Escher, Bach (GEB)* are a mixture of his musings on consciousness, intelligence (human and artificial), mathematics and a whole lot more. GEB won the Pulitzer Prize for General Non-Fiction in 1980. The Strange Loop Conferences, sadly, ended in 2023.

Textbooks on Big Data Programming

Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale, 4th Edition Tom White	O'Reilly Media Copyright: 2015 ISBN: 978-1491901632
Spark: The Definitive Guide Bill Chambers Matei Zaharia	O'Reilly Media Copyright: 2018 ISBN: 978-1491912218

Grades for the Course

Item	% score
Class Participation Your class participation score is intended to reflect your effort in maintaining a collaborative learning environment for everyone. ● Turning on your video camera and keeping it turned on!, ● Thoughtful (and helpful) questions/comments in class and on Piazza, ● Willingness to help peers when they are stuck, (without doing the work for them), ● Participation in office hours.	**8%**
Academic paper reviews Big Data is a continuously evolving field. The ability to consume published, peer-reviewed papers is critical your ability to stay current with it. We will follow a methodology proposed by S. Keshav to review some of the seminal papers in the field.	**12%**
Quizzes The quizzes are intended to reflect your effort in staying current with what's happening in class. They are typically due a week after they are assigned, never more than two weeks. Only the best 70% (7 out of 10) of your quiz scores will count towards the final grade.	**60%**

Gedanken Project[3] Writing proposals is an integral part of moving up in the organization where you work, be it academe or industry. Most significant Research projects involve putting together a system, obtaining the data and running experiments. The first step in the research is to garner funding for it. You will not be doing the actual research just yet — you will be creating a *proposal*[4] for securing funding for the research. Students may choose from 6 project ideas (to be announced during week 5) or propose their own.	**20%**

Course Administration and Policies

Lab Environment

Each student will have an allowance for use of a Cloud Platform to be used for Quizzes 2-6. Google has been generous in making this resource available to us through their "Google for Education" program, please be thoughtful in using it. Some things to consider:

- *Shut off your VM or cluster when you are done using it, otherwise they will continue to accrue charges.*
- *Don't confuse Google's "Free Trial" program with the "Google for Education" program* and be aware of the risk of signing up for the Free Trial. It is deceptively easy to change to billing that credit card when the $300/90 days of the Free Trial is used up.
- *If your credit card is charged at the end of the Free Trial program, Google for Education cannot help.* Students are asked before changing to credit card billing, but they often just click the button to upgrade their account without understanding what that means. Please see this warning from Google for Education. Additionally, the Free Trial program has a few more restrictions than the Google for Education program and can only be used once.

[3] "Gedanken Project" is named after Gedanken Experiment.
[4] A research proposal is a detailed plan or 'blueprint' for the intended study, and once it is completed, the research project should flow smoothly. The Gedanken Project will include the technical design of a system but not its implementation.

Seeking Help

Please keep in mind the following as regards to approaching the instructor for help.

- For technical questions, please utilize Piazza — plus other students could likely benefit from the Q & A. Some students may answer your questions even faster! *Public questions will receive a higher priority because they apply to everyone.*
- For questions that have specifics of your solution that you don't want others to see, it's OK to post private questions.
- To get in touch with the instructor for a matter unrelated to course content, please use email. Please keep the use of email to confidential matters, not for general class discussion
- If there is no response from the above within 24 hrs, or in case of an emergency, please call the instructor.

Illness-related policies

Please do not come to class when exhibiting even mild Covid-19 symptoms. If you are ill or symptomatic, please alert the teaching team via Piazza. This guideline applies to everyone: the students, TAs and the instructor.

Zoom recordings will be available to those who are unable to attend due to illness, anxiety, grief or trauma. More on illness-related policies below.

To make zoom recordings available for your unexpected absences, please be aware that you are consenting to being recorded. If you have objections to being recorded, please contact me before class.

If unable to attend the lecture in person, please consider attending Synchronously over Zoom. You will need to sign in using your tufts credentials.

Academic Integrity

You are expected to be familiar with the Student Guide to Academic Integrity at Tufts and follow those guidelines.

Accommodations for Students with Disabilities

Tufts University values the diversity of our students, staff, and faculty and recognizes the important contribution each student makes to our unique community. Tufts is committed to providing equal access and support to all qualified students through the provision of reasonable accommodations so that each student may fully participate in the Tufts experience. If you have a disability that requires reasonable accommodations, please contact the StAAR Center (formerly Student Accessibility Services) at StaarCenter@tufts.edu or 617-627-4539 to make an appointment with an accessibility representative to determine appropriate accommodations. Please be aware that *accommodations cannot be enacted retroactively*, making timeliness a critical aspect for their provision.

Academic Support at the StAAR Center

The StAAR Center (formerly the Academic Resource Center and Student Accessibility Services) offers a variety of resources to all students (both undergraduate and graduate) in the Schools of Arts and Science, Engineering, the SMFA and Fletcher; services are free to all enrolled students. Students may make an appointment to work on any writing-related project or assignment, attend subject tutoring in a variety of disciplines, or meet with an academic coach to hone fundamental academic skills like time management or overcoming procrastination. Students can make an appointment for any of these services by visiting the StAAR Center website.

Mental Health Support

As a student, there may be times when personal stressors or emotional difficulties interfere with your academic performance or well-being. The Counseling and Mental Health Service (CMHS) provides confidential consultation, brief counseling, and urgent care at no cost for all Tufts undergraduates as well as for graduate students who have paid the student health fee. To make an appointment, call 617-627-3360. Please visit the CMHS website to learn more about their services and resources.

About the Instructor

I have worked in Cloud Computing, Big Data and Python since 2008. Python has been my programming language of choice ever since!

I received my Ph.D. in Electrical Engineering working on solving large-scale matrix problems in Electromagnetics. I was initially on the EE faculty at WPI, then left to spend a major part of my career in industry, mostly in Systems Architect roles, first Computer-Aided Design and later in Finance. Throughout my career, I have stayed close to data and databases as my area of focus. I rejoined WPI in the CS department in 2009 teaching Databases and Big Data. I've been with Tufts since 2018.

Please call me Jitendra or J or Prof. J, whichever you prefer. (No period after the J)