Syllabus

CSCI 493.71: Big Data

Instructor: Lei Xie | **Office:** online
Email: lxie@iscb.org | **Telephone:** 212-396-6550
Office hours: by appoitment
Lecture time: Monday, 5:35pm -7:35pm
Classroom: online

Topics, Goals or Outcomes

This course will introduce fundamental knowledge and skills on *big data* processing by using cloud computing technology. The application of big data analytics in biomedicine will be used as primary examples. Thus, this course is intended for both students in Quantitative Biology program and computer science majors.

The increasing availability of big data has changed fundamental practices in business, health care, policy making, and scientific research. Data scientist will be one of the most demanded jobs of the 21st century (Harvard Business Review, Oct. 2012). **The primary objective of this course is to enable you, working under a UNIX/LINUX and Amazon EC2 environment, to process data, manage data, extract information, and discover patterns from big data.** Although we will often work with biological data, the fundamental principles you will learn can be applied in any domain.

This course is organized around three essential questions:

- What is data science and data engineering?
- What are new paradigms of big data?
- What are pros and cons of existing big data technologies?
- How can we extract information and discover knowledge from big data?

Students' responsibilities

Students are expected to attend all classes on time and to be well prepared. Attendance will be taken at each class meeting. All electronic devices (including cell phones) are to be turned off during class.

Students are expected to maintain a UNIX account with the Department and abide by the rules for access and use of the Department's laboratory facility.

Students are responsible for all material in the reading, whether or not it is covered during class time. Classes will presume that you have read the assigned material. Come to class prepared to ask, and to answer, questions on the reading. The text is a necessary resource, but lectures will go far beyond it. Detailed notes are highly recommended, and questions during class are encouraged. If you miss a class, it is your responsibility to get notes, assignments, and prepare for tests.

Students are responsible for all email sent to their registered Hunter Blackboard address. You should forward your Hunter mail to an address you read regularly. Changes in assignments, clarifications, and instructions will often be distributed by email.

Students must acknowledge any help received from other individuals and must reference in full any material used (e.g., encyclopedia, book, paper, journal, web site) to prepare assignments.

Students are expected to submit all projects at the *beginning* of class on its due date. For example, an assignment due on a Tuesday is due at 2:10 PM on Tuesday, and must be delivered to the instructor in class. If you must miss class (never a good idea), get your homework *to me* *before class*. Otherwise, you must have your homework *time-stamped by the Department office* HN 1008 *and left in my mailbox there*. (Do *not* try to put it under my office door.)

Textbook (Recommended but not Required)

- Big Data Science & Analytics, A Hands-On Approach, by Arshdeep Bahga, Vijay Madisetti

Software

You need install several software packages on your own computer, and register an Amazon EC2 account.

Grading and project

- **Course grades** are based on three projects, and thoughtful, well-prepared class participation:
- 10% class participation (asking questions counts!)
- 30% quiz
- 60% project

Hunter College Policy on Academic Integrity

“Hunter College regards acts of academic dishonesty (e.g., plagiarism, cheating on examinations, obtaining unfair advantage, and falsification of records and official documents) as serious offenses against the values of intellectual honesty. The College is committed to enforcing the CUNY Policy on Academic Integrity and will pursue cases of academic dishonesty according to the Hunter College Academic Integrity Procedures.”

ADA compliance

In compliance with the American Disability Act of 1990 (ADA) and with Section 504 of the Rehabilitation Act of 1973, Hunter College is committed to ensuring educational parity and accommodations for all students with documented disabilities and or/or medical conditions. It is recommended that all students with documented disabilities (Emotional, Medical, Physical and/or Learning) consult the Office of AccessABILITY located in Room E1124 to secure necessary academic accommodations. For further information and assistance please call (212-772-4857)/TTY (212-650-3230).

Hunter College Policy on Sexual Misconduct

“In compliance with the CUNY Policy on Sexual Misconduct, Hunter College reaffirms the prohibition of any sexual misconduct, which includes sexual violence, sexual harassment, and gender-based harassment retaliation against students, employees, or visitors, as well as certain intimate relationships. Students who have experienced any form of sexual violence on or off campus (including CUNY-sponsored trips and events) are entitled to the rights outlined in the Bill of Rights for Hunter College.

a. Sexual Violence: Students are strongly encouraged to immediately report the incident by calling 911, contacting NYPD Special Victims Division Hotline (646-610-7272) or their local police precinct, or contacting the College's Public Safety Office (212-772-4444).

b. All Other Forms of Sexual Misconduct: Students are also encouraged to contact the College's Title IX Campus Coordinator, Dean John Rose (jtrose@hunter.cuny.edu or 212-650-3262) or

Colleen Barry (colleen.barry@hunter.cuny.edu or 212-772-4534) and seek complimentary services through the Counseling and Wellness Services Office, Hunter East 1123.

CUNY Policy on Sexual Misconduct Link: *http://www.cuny.edu/about/administration/offices/la/Policy-on-Sexual-Misconduct-12-1-14-with-links.pdf*

Electronic device usage policy

I expect all cell phones, pagers, etc. to be inaudible during class. I expect laptops and other electronic devices, if used, to be used only for class related activities. Activities not related to class include but are not limited to facebook, twitter, other social networking web sites, “surfing", email, mu*s, hulu, southparkstudios, etc. Any student with an electronic device that disrupts the class will lose two (2) points from their final average (per occurrence).

An evolving schedule

The schedule here may evolve as the semester progresses.

Topic	**Date**	**Note**
Introduction to big data	8/31	
Relational database system & data warehouse	9/14	
NoSQL: Fundamental & Document Store	9/21	Q1
NoSQL: Key-value & Column Family	9/29	Q2
NoSQL: Neo4j & Cypher	10/5	**Q3** Project I assignment
Hadoop & HDFS	10/14	Q4
Spark	10/19	Q5
MapReduce (I): Introduction	10/26	Q6
Project II review	11/2	
MapReduce (II): Design pattern	11/9	Q7 Project II assignment

Streaming algorithm	11/16	Q8
Informational retrieval	11/23	Q9
Data mining & machine learning (I)	11/30	Q10 https://towardsdatascience.com/i-read-one-
Data mining & machine learning (II)	12/7	
Final project review	12/16	