New York University Tandon School of Engineering
Computer Science and Engineering
CS-GY 6513 A Big Data
Fall 2019
Professor Julia Stoyanovich
Time: Monday 12:25-2:55pm
Location: 2 MetroTech Center, Room 9.011

To contact professor: stoyanovich@nyu.edu
370 Jay Street, Room 1101
Office hours:Tuesday 10am-noon or by appointment

Course Prerequisites: Familiarity with the Python programming language.

Course Description: Big Data requires the storage, organization, and processing of data at a scale and efficiency that go well beyond the capabilities of conventional information technologies. In this course, we will study the state of the art in big data management: we will learn about algorithms, techniques and tools needed to support big data processing. In addition, we will examine real applications that require massive data analysis and how they can be implemented on Big Data platforms. The course will consist of lectures based both on textbook material and scientific papers. It will include programming assignments that will provide students with hands-on experience on building data-intensive applications using existing Big Data platforms. Besides lectures given by the instructor, we will also have guest lectures by experts in some of the topics we will cover.

Course Objectives: After successfully completing the course, students are able to:

- Explain the characteristics of Big Data applications, and articulate the main design principles of data processing platforms that respond to these characteristics.
- Use MapReduce and Apache Spark to efficiently implement large-scale data analysis of urban transportation datasets.
- Articulate the need for data profiling and data cleaning, and implement several data profiling and data cleaning methods using open-source tools and libraries.
- Implement a computer program that efficiently identifies frequently occurring patterns in large datasets.
- Articulate the difference between different interpretations of algorithmic fairness, and implement a computer program that demonstrates the trade-offs.

Course Structure: The course is structured into a sequence of lectures, labs, and accompanying assignments. The assignment consists of three homeworks and a course project.

- Labs are short exercises done in class and submitted in class.

- Homeworks are longer exercises designed to take two weeks to complete. Homeworks are to be completed individually.
- The course project is designed to take 10 weeks to complete. Projects are to be completed in teams of three.

Readings: This course does not have a required textbook. Each topic will be accompanied by required reading, as listed in the weekly schedule. In some cases, expert-level technical research papers are listed as assigned reading. Some readings will be chapters from *Mining of Massive Datasets*, 2nd edition, available online at http://www.mmds.org/, and of research papers.

Course Assessment

Students accumulate up to 100 points during the course.

- **Homeworks**: 3 x 10 points per homework = 30 points. Homeworks are assigned on Monday in class, and are due in two weeks, at 9am on a Monday. Homeworks must be submitted on time. If an assignment is submitted late, the student will receive no credit.
- **Midterm exam**: 30 points. The exam will be conducted in class.
- **Project**: 30 points.
- **Attendance and participation:** 10 points. Attend 10 lectures for full credit. October 28 (midterm exam) and December 16 (final project presentation) do not count towards the 10-lecture attendance requirement.

Grades will be determined using this scale:

Course Grade	Points Earned
A	94-100
A-	90-93
B+	87-89
B	84-86
B-	80-83
C+	76-79
C	72-75
C-	70-71
D+	66-69
D	62-65
D-	60-61
F	less than 60

Course requirements

Students are expected to:

- attend lectures and labs
- complete assigned readings before class
- submit homework assignments and the course project on time
- take an in-class midterm exam

Homework 1:

- Assigned 09/16/2019, due 09/30/2019
- 10% of the course grade

Homework 2:

- Assigned 10/07/2019, due 10/21/2019
- 10% of the course grade

Homework 3:

- Assigned 11/11/2019, due 11/25/2019
- 10% of the course grade

Midterm exam:

- 10/28/2019, in class
- 30% of the grade

Project:

- Assigned 09/23/2019, due 12/12/2019, presentation 09/16/2019
- 30% of the course grade

Course Schedule

09/09/2019 1. Lecture + Lab: Course Introduction, Map Reduce

Reading:

- Big data and its technical challenges, Jagadish et al., Comm ACM 57(7), 2014
- Mining of Massive Datasets, 2nd ed, Ullman & Leskovec, Chapter 2
- Data-Intensive Text Processing with MapReduce, Lin & Dyer, Chapters 2, 3.1

Assignments: none

09/16/2019 2. Lecture + Lab: Apache Spark

Reading:

- Apache Spark: a unified engine for big data processing, Zaharia et al., Comm ACM 59(11), 2016
- Resilient Distributed Datasets: A fault-tolerant abstraction for in-memory cluster computing, Zaharia et al., NSDI 2012

Assignments: HW1 assigned

09/23/2019 3. Lecture: The relational model, SQL, Lab: Spark SQL

Reading:

- Spark SQL: Relational data processing in Spark, Armbrust et al., ACM SIGMOD 2015.

Assignments: course project announced, work on HW1

09/30/2019 4. Lecture + Lab: Spatial analytics

Reading: TBD

Assignments: HW1 due at 9am, work on the project

10/07/2019 5. Lecture + Lab: Frequent itemset mining

Reading:

- Fast algorithms for mining association rules, Agrawal & Srikant, VLDB 1994
- Mining of Massive Datasets, 2nd ed, Ullman & Leskovec, Chapter 6

Assignments: HW2 assigned

10/14/2019 6. Lecture + Lab: Data profiling, data cleaning

Reading:

- Profiling relational data: A survey, Abedjan et al, VLDB Journal (24), 2015
- Quantitative data cleaning for large databases, Hellerstein, UNECE 2008

Assignments: work on HW2, project

10/21/2019 7. Lecture: Similarity, Lab: Midterm review

Reading:

- Mining of Massive Datasets, 2nd ed, Ullman & Leskovec, Chapter 3

Assignments: HW2 due at 9am, prepare for midterm exam

10/28/2019 8. Lecture + Lab: midterm exam, in class

Reading: none

Assignments: work on the project

11/04/2019 9. Lecture + Lab: Open Data & Data Lakes

Reading: TBD

Assignments: work on the project

11/11/2019 10. Lecture: Algorithmic fairness, Lab: FairBench

Reading:

- Machine bias, Angwin et al., ProPublica 2016
- Inherent trade-offs in the fair determination of risk scores, Kleinberg et al., ITCS 2017
- On the (im)possibility of fairness, Friedler et al., arXiv 2016

Assignments: HW3 assigned

11/18/2019 11. Lecture: Urban analytics, Lab: recent research on urban analytics

Reading:

- SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution, Bello et al., Comm ACM 62(2), 2019

Assignments: work on HW3, project

11/25/2019 12. Lecture + Lab: Reproducibility

Reading: TBD

Assignments: HW3 due at 9am, work on the project

12/02/2019 13. Lecture: Transparency and interpretability, Lab: LIME

Reading:

- Why should I trust you? Explaining the predictions of any classifier, Ribeiro et al., ACM KDD 2016
- Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes, Ali et al., arXiv 2019

Assignments: work on the project

12/09/2019 14. Lecture: Recommender systems, Lab: recent research

Reading: TBD

Assignments: work on the project

12/16/2019 15. Course project presentations

Moses Center Statement of Disability

If you are a student with a disability who is requesting accommodations, please contact New York University's Moses Center for Students with Disabilities (CSD) at 212-998-4980 or mosescsd@nyu.edu. You must be registered with CSD to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csd. The Moses Center is located at 726 Broadway on the 3rd floor.

NYU School of Engineering Policies and Procedures on Academic Misconduct – complete Student Code of Conduct here

A. Introduction: The School of Engineering encourages academic excellence in an environment that promotes honesty, integrity, and fairness, and students at the School of Engineering are expected to exhibit those qualities in their academic work. It is through the process of submitting their own work and receiving honest feedback on that work that students may progress academically. Any act of academic dishonesty is seen as an attack upon the School and will not be tolerated. Furthermore, those who breach the School's rules on academic integrity will be sanctioned under this Policy. Students are responsible for familiarizing themselves with the School's Policy on Academic Misconduct.

B. Definition: Academic dishonesty may include misrepresentation, deception, dishonesty, or any act of falsification committed by a student to influence a grade or other academic evaluation. Academic dishonesty also includes intentionally damaging the academic work of others or assisting other students in acts of dishonesty. Common examples of academically dishonest behavior include, but are not limited to, the following:

1. Cheating: intentionally using or attempting to use unauthorized notes, books, electronic media, or electronic communications in an exam; talking with fellow students or looking at another person's work during an exam; submitting work prepared in advance for an in-class examination; having someone take an exam for you or taking an exam for someone else; violating other rules governing the administration of examinations.
2. Fabrication: including but not limited to, falsifying experimental data and/or citations.
3. Plagiarism: intentionally or knowingly representing the words or ideas of another as one's own in any academic exercise; failure to attribute direct quotations, paraphrases, or borrowed facts or information.
4. Unauthorized collaboration: working together on work meant to be done individually.
5. Duplicating work: presenting for grading the same work for more than one project or in more than one class, unless express and prior permission has been received from the course instructor(s) or research adviser involved.
6. Forgery: altering any academic document, including, but not limited to, academic records, admissions materials, or medical excuses.

NYU School of Engineering Policies and Procedures on Excused Absences – complete policy here

A. Introduction: An absence can be excused if you have missed no more than **10 days of school**. If an illness or special circumstance has caused you to miss more than two weeks of school, please refer to the section labeled Medical Leave of Absence.

B. Students may request special accommodations for an absence to be excused in the following cases:

1. Medical reasons
2. Death in immediate family
3. Personal qualified emergencies (documentation must be provided)
4. Religious Expression or Practice

Deanna Rayment, deanna.rayment@nyu.edu, is the *Coordinator of Student Advocacy, Compliance and Student Affairs* and handles excused absences. She is located in 5 MTC, LC240C and can assist you should it become necessary.

NYU School of Engineering Academic Calendar – complete list here.
The last day of the final exam period is 12/20/2019. Final exam dates for undergraduate courses will not be determined until later in the semester. Final exams for graduate courses will be held on the last day of class during the week of 12/16/2019. If you have two final exams at the same time, report the conflict to your professors as soon as possible. Do not make any travel plans until the exam schedule is finalized.

Also, please pay attention to notable dates such as Add/Drop, Withdrawal, etc. For confirmation of dates or further information, please contact Susana: sgarcia@nyu.edu