Boston University Metropolitan College

Big Data Analytics

MET CS777

Farshid Alizadeh-Shabdiz, PhD, MBA
alizadeh@bu.edu
Office hours: by appointment

Course Description

This course is an introduction to large-scale data analytics. Big Data analytics is the study of how to extract actionable, non-trivial knowledge from massive amount of data sets. This class will focus both on the cluster computing software tools and programming techniques used by data scientists, as well as the important mathematical and statistical models that are used in learning from large-scale data processing. On the tools side, we will cover the basics systems and techniques to store large-volumes of data, as well as modern systems for cluster computing based on Map-Reduce pattern such as Hadoop MapReduce and Apache Spark.

Students will implement data mining algorithms and execute them on real cloud systems like Amazon AWS, Google Cloud or Microsoft Azure by using educational accounts. On the data mining models side, this course will cover the main standard supervised and unsupervised models and will introduce improvement techniques on the model side.

Prerequisites
MET CS 521 and either MET CS 622, MET CS 673 or MET CS 682.
MET CS 677 is strongly recommended.
Or, instructor's consent.

Learning Objectives

By successfully completing this course you will be able to:

- Learn the main challenges of Big Data Processing
- Run a Big Data Processing pipeline on Google Cloud Implement Big Data
- Coding Apache Spark in PySpark
- Run Supervised and Unsupervised machine learning on Large-Scale Data

Boston University Metropolitan College

BOSTON UNIVERSITY

Required Text Book

There is no required textbook for the class. All class material will be conveyed during lecture.

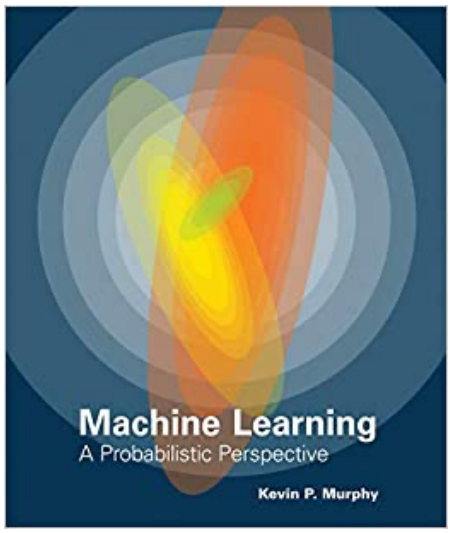
Recommended Books and Material

Strongly suggest using one or two of the books as a complement.

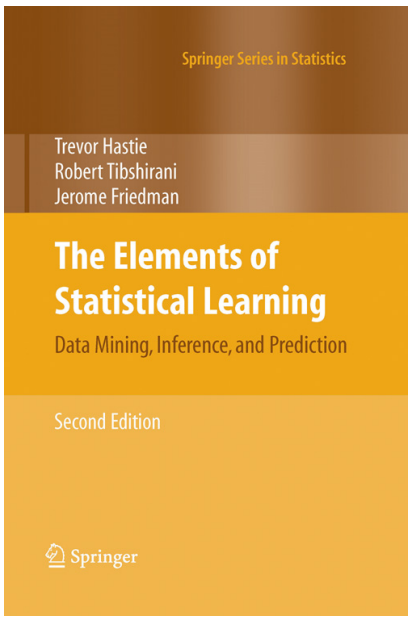
1. Main Apache Spark documentation website

   https://spark.apache.org/docs/latest/

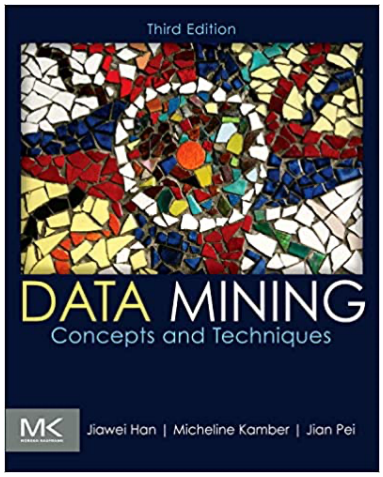
1) **Murphy, K. (2012).** ***Machine learning: a probabilistic perspective*** The MIT Press
ISBN-13: 978-0262018029



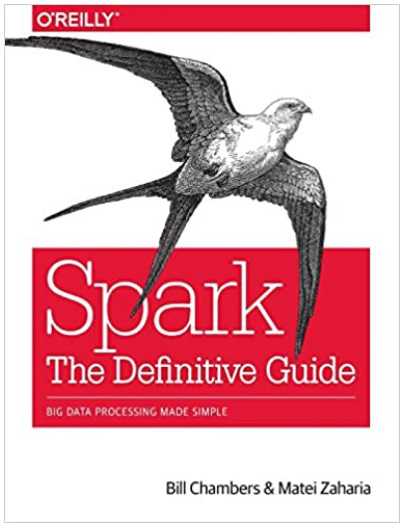
2) **Hastie, T. and Tibshirani, R. (2009).** ***The elements of statistical learning: Data mining, inference, and prediction*** **(2nd ed.).**
Springer-Verlag.
ISBN-13: 978-0-387-84858-7
This book is available online in PDF format.



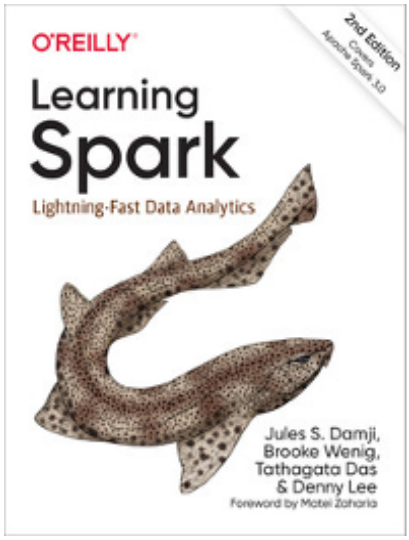
3) **Han, J., Kamber, M., Pei, J. (2009).** ***Data mining: Concepts and techniques*** **(3rd ed.).** Morgan Kaufmann. ISBN-13: 978-9380931913



4) Spark The Definitive Guide, Chambers, B. & Zaharia, M. (2018). *Spark: The definitive guide: Big data processing made simple* O'Reilly Media Inc.



5) **Learning Spark, 2nd edition.** Damji, J., Wenig, B., Das, T., Lee, D. (2020). *Learning spark* (2nd ed.) O'Reilly Media Inc.



Cloud System

In this class, real-world cloud systems, Google Cloud and AWS will be used as examples. Students will receive educational credit coupons to the cloud systems.

You should never use your private account or use your credit card for this class assignment. You will receive enough education credits so you can run successful assignments.

The credit amount is $50 USD for Google Cloud. You should use only this amount to finish your assignments. This will be more than enough to finish the assignments, learn Google Cloud work, and have your first enjoyable experience with it. The credit amount given to students by Amazon to use AWS is $100.

This is real money. Therefore, students have to develop your code and run your jobs locally, your laptop, using the small data set. Once things are working, you'll then move to the Cloud. We will ask you to run your Spark jobs over the "real" data using a set of cluster machines.

Courseware

List course website (Blackboard), as well as any web links that will be necessary for the class.

COVID-19 Policies

Compliance: All students returning to campus will be required, through a digital agreement, to commit to a set of Health Commitments and Expectations including face coverings, testing, contact tracing, quarantine, and isolation. The agreement makes clear that compliance is a condition of being a member of our on-campus community.

You have a critical role to play in minimizing transmission of COVID-19 within the University community, so the University is requiring that you make your own health and safety commitments. Additionally, you are asked to always wear your face mask over your mouth and nose. If you do not comply with these rules you will be asked to leave the classroom. If you refuse to leave the class, the instructor will inform the class that they will not proceed with instruction until you leave the room. If you still refuse to leave the room, the instructor will dismiss the class and will contact the academic Dean's office for follow up.

Boston University is committed to offering the best learning environment for you, but to succeed, we need your help. We all must be responsible and respectful.

Class Policies

1) **Assignment Completion & Late Work** – all the assignment has to be submitted on time on Blackboard. No late work will be acceptable.
2) **Academic Conduct Code** – Cheating and plagiarism will not be tolerated in any Metropolitan College course. They will result in no credit for the assignment or examination and may lead to disciplinary actions. Please take the time to review the Student Academic Conduct Code:

http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/code.html.

Grading Criteria

The course grade will be based on

- Assignments - 35%
- Take home and in-class quizzes and possible final exam - 20%
    - If there is a final exam, its weight will be 2x a quiz
- Term paper - individually - 15%
    - Students individually choose a subject, relevant paper or a technology related to the subject of the course
    - Students individually prepare a report
        - Maximum one page executive summary
        - Five to ten page report
        - Maximum 3 pages of supporting code
    - Students prepare individual 10-min presentation
    - Although the work is individually, students can team up with maximum two other students and choose related technologies. In case of working as a team, 80% of the grade comes from individual work and 20% from the team work.

- Term project – 30%
    - Team work with the team of 2 or 3 people
    - Term project must consist of processing a large data (multi-GB) and coding
    - Deliverables are project report, copy of the code, 10 min presentation in the class, and copy of the presentation slides

Tentative Class Syllabus

Lectures, Readings, and Assignments subject to change, and will be announced in class as applicable within a reasonable time frame.

Tentative Class Schedule

Lectures, Readings, and Assignments subject to change, and will be announced in class as applicable within a reasonable time frame.

Module on Map Reduce and Spark Data Processing Pattern

- Introduction to Big Data Analytics.

- What is Big Data? What are the challenges?
- Introduction to Apache Hadoop and MapReduce.
- Apache Spark.
- Spark programming - Python and pySpark.
- Spark Resilient Distributed Dataset (RDDs).

Module on Large-Scale Data Processing and Storage

- Spark DataFrames
- Spark SQL
- Code Optimization and Cluster Configurations
- Recap – Relational Databases and SQL
- Distributed Object Storage Systems

Module on PySpark+Numpy and Scipy in Big Data

Module on big data optimization

Module on Introduction to Modeling and Optimization Basics

- Optimization basics: Gradient descent (batch and stochastic),
- Selected machine learning topics
- Generalized linear model
- Newton Method

Module on Machine Learning on Large-Scale Data

- Spark MLlib
- Spark ML
- Deployment model of ML algorithms

Module on Streaming

- Spark streaming