Syllabus
MET CS 777 O1 Big Data Analytics 2024 Spring 1
Classes: Tuesdays 5:30-7:00 pm and Thursdays 5:30-7:00 pm
(from 1/16/24 to 3/4/24)

Description

MET CS 777 Big Data Analytics

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 a massive number of data sets. This class will focus both on the cluster computing software tools and programming techniques used by data scientists and the important mathematical and statistical models used in learning from large-scale data processing. On the tool's side, we will cover the basic systems and techniques to store large volumes of data and modern systems for cluster computing based on MapReduce patterns such as Hadoop MapReduce and Apache Spark.

Students will implement data mining algorithms and execute them on real cloud systems like Google Cloud, Amazon AWS, 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.

Course Prerequisites

- We expect you to have a solid background in Python programming and understand basic statistics and machine learning. The following classes are required/recommended: MET CS 521, MET CS 544 and MET CS 555, or MET CS 677.
- If you do not have the required/recommended courses, you need the instructor's consent.
- This class includes topics from Cloud Computing, Parallel Processing, and Machine Learning, which make the course very compact for a six-week online course.
- To implement the assignments, students need to have excellent knowledge of Python programming language and some basic Linux knowledge. Assignments are very time-consuming, and you should take this course when you have at least 20 hours per week.

Learning Objectives

By successfully completing this course, you will be able to:

- Explain the main challenges of Big Data Processing
- Run a Big Data Processing pipeline on Google Cloud (or Amazon AWS)
- Implement Big Data code in Apache Spark (in PySpark)
- Run Supervised and Unsupervised machine learning on Large-Scale Data

Instructional Team

Instructor: Dimitar Trajanov, PhD

Computer Science Department
Metropolitan College
Boston University

email: dtrajano@bu.edu

Prof. Dimitar Trajanov, Ph.D., is a visiting research professor at Boston University and a full professor at the Faculty of Computer Science and Engineering. Cyril and Methodius University—Skopje. From March 2011 until September 2015, he was the founding dean of the Faculty of Computer Science and Engineering, and during his tenure, the faculty became the largest technical faculty in Macedonia. Dimitar Trajanov is the leader of the Regional Social Innovation Hub, which was established in 2013 as a cooperation between UNDP and the Faculty of Computer Science and Engineering.

His professional experience includes working as a Senior Data Science Consultant for one of the largest Pharmaceutical companies, a Data Science consultant for UNDP in North Macedonia, and a software architect in a couple of startups.

Dimitar Trajanov is the author of more than 200 journal and conference papers and seven books. He has been involved in more than 80 research and industry projects.

Materials

Required Book

There is no required textbook for the class. All class material will be conveyed during the lecture. The following recommended books and materials are available online.

Materials and Digital Learning Assets

Online Materials

The course adopts a "learning by example" approach, offering a hands-on approach through more than 50 Python notebooks and scripts. This educational strategy fosters real-world relevance, deepens theoretical understanding, and hones both hard and soft skills essential to the field.

Please download the CS-777 Learning by Example Guide (PDF).

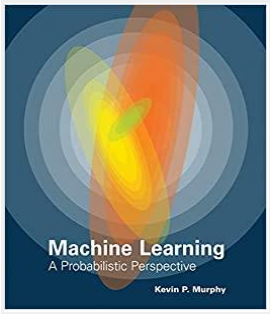
Spark Online Guides

The official PySpark documentation is maintained by the developers themselves, ensuring that it is always up-to-date and accurate with the latest changes in the platform.

The following guides are highly recommended for anyone learning or working with PySpark:

- PySpark Getting Started: This is the Getting Started guide that summarizes the basic steps required to setup and get started with PySpark.
- RDD Programming Guide: This guide provides an overview of Spark basics, including RDDs (the core but older API), accumulators, and broadcast variables.
- Spark SQL, Datasets, and DataFrames: This guide is focused on processing structured data with relational queries using a newer API than RDDs.
- MLlib: This guide provides detailed information on how to apply machine learning algorithms in PySpark.
- Spark Python API. This is the PySpark API documentation, which provides detailed information on the PySpark API, including its modules, classes, and methods.

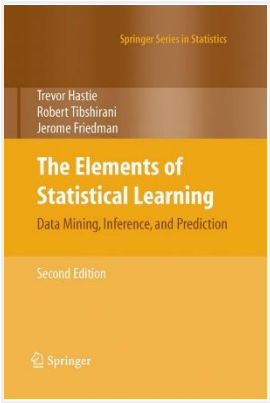
Books



Murphy, K. (2012). Machine learning: a probabilistic perspective

The MIT Press

ISBN-13: 978-0262018029

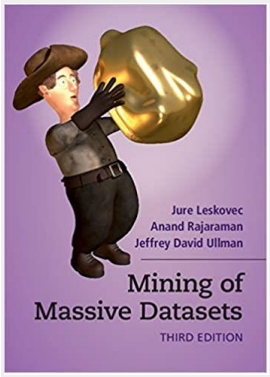


Hastie, T. and Tibshirani, R. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.).

Springer-Verlag.

ISBN-13: 978-0-387-84858-7

This book is available for PDF download.

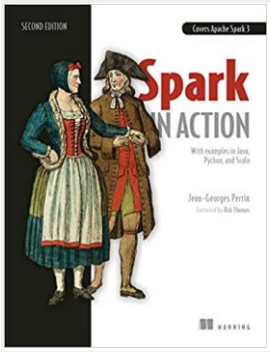
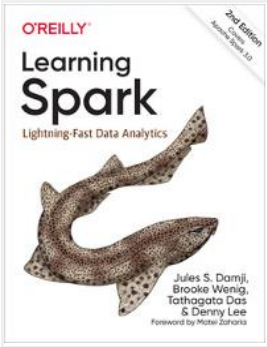
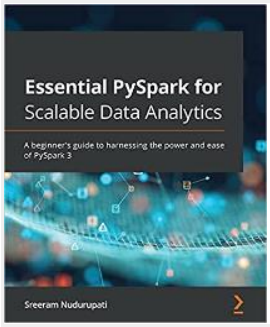
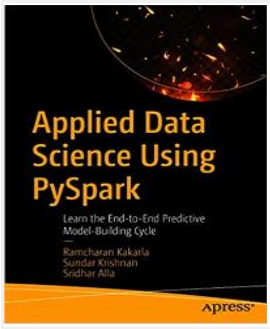


Leskovec, J. Rajaraman, A., Ullman, J. (2014). Mining of massive datasets.

Cambridge University Press.

By agreement with the publisher, you can download the book for free from this page

Other Materials and Resources

	Perrin, J. (2020). Spark in action (2nd ed.). (Covers Apache Spark 3 with examples in Java, Python, and Scala) O'Reilly Media Inc.
	Damji, J., Wenig, B., Das, T., Lee, D. (2020). Learning spark (2nd ed.) O'Reilly Media Inc.
	Nudurupati, S. (2021). Essential PySpark for scalable data analytics: A beginner's guide to harnessing the power and ease of PySpark 3 Packt Publishing
	Ramcharan, K., Sundar, K., Alla, S. (2020). Applied data science using PySpark: Learn the end-to-end predictive model-building cycle Apress
	Main Apache Spark documentation website

GitHub

This course has a **GitHub repository** with all of the course code examples.

Usage of Cloud Machines

In this class, we use real-world cloud systems existing on Google Cloud (or Amazon AWS). You will receive educational credit coupons or credited access to such cloud systems. You should never use your private account or use your credit card for this class assignment. You will receive enough education credits to run successful assignments on Google Cloud.

The credit amount is 50 USD for Google Cloud. You should use only this amount to finish your assignments. This would be more than enough to finish the assignments, learn how Google Cloud (or AWS) works, and have your first enjoyable experience with it. You can choose different numbers of Machines and different configurations of those machines. And each will cost you differently!

Since this is real money, it makes sense to develop your code and run your jobs locally, on your laptop, using the small data set (we will provide two types of the same data set, small and big). Once things are working, you'll then move to Amazon AWS or Google Cloud. We will ask you to run your Spark jobs over the "real" data using a set of cluster machines.

Study Guide

This course starts on a **Tuesday**. The modules in this course run from **Tuesday to Monday**.

Module 1 Study Guide and Deliverables
MapReduce Data Processing Pattern

Readings:

- Online lecture material topics:
  - 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)

Assignments:

- Assignment 1 due Saturday, January 27 at 6:00 AM ET

Assessments:

- Quiz 1 due Friday, January 26 at 6:00 AM ET

Live Classroom:

- Tuesday, January 16 from 5:30–7:00 PM ET
- Thursday, January 18, from 5:30–7:00 PM ET
- Live Office (facilitator sessions): to be scheduled each weekend.

Module 2 Study Guide and Deliverables
Large-Scale Data Processing and Storage

Readings:

- Online lecture material topics:
  - Spark - RDDs, DataFrames, Spark SQL
  - PySpark + NumPy
  - Code Optimization, Cluster Configurations
  - Distributed File Storage Systems

Assignments:

- Assignment 2 due Wednesday, January 31 at 6:00 AM ET

Assessments:

- Quiz 2 due Tuesday, January 30 at 6:00 AM ET

Live Classroom:

- Tuesday, January 23, from 5:30–7:00 PM ET
- Thursday, January 25 from 5:30–7:00 PM ET
- Live Office (facilitator sessions): to be scheduled each weekend

Module 3 Study Guide and Deliverables
Data Modeling and Optimization Problems

Readings:

- Online lecture material topics:
  - Introduction to Modeling: Numerical vs. Probabilistic vs. Bayesian
  - Introduction to Optimization Problems
  - Batch and Stochastic Gradient Descent
  - Newton's Method
  - Expectation Maximization

Assignments:

- Assignment 3 due Wednesday, February 7 at 6:00 AM ET

Assessments:

- Quiz 3 due Tuesday, February 6 at 6:00 AM ET

Live Classroom:

- Tuesday, January 30 from 5:30–7:00 PM ET
- Thursday, February 1 from 5:30–7:00 PM ET
- Live Office (facilitator sessions): to be scheduled each weekend.

Module 4 Study Guide and Deliverables
Large-Scale Supervised Learning

Readings:

- Online lecture material topics:
  - Introduction to Supervised Learning
  - Generalized Linear Models and Logistic Regression
  - Regularization
  - Spark ML library

Assignments:

- Assignment 4 due Wednesday, February 14 at 6:00 AM ET

Assessments:

- Quiz 4 due Tuesday, February 13 at 6:00 AM ET

Live Classroom:

- Tuesday, February 6 from 5:30–7:00 PM ET
- Thursday, February 8, from 5:30–7:00 PM ET
- Live Office (facilitator sessions): to be scheduled each weekend.

Module 5 Study Guide and Deliverables
Unsupervised Learning on Large-Scale Data

Readings:

- Online lecture material topics:
  - Introduction to Unsupervised learning
  - K-means / K-medoids
  - Gaussian Mixture Models
  - Matrix factorization
  - Dimensionality Reduction

Assignments:

- Term Project Proposal due Tuesday, February 20 at 6:00 AM ET
- Assignment 5 due Wednesday, February 21 at 6:00 AM ET

Assessments:

- Quiz 5 due Tuesday, February 20 at 6:00 AM ET

Live Classroom:

- Tuesday, February 13 from 5:30–7:00 PM ET
- Thursday, February 15 from 5:30–7:00 PM ET
- Live Office (facilitator sessions): to be scheduled each weekend.

Module 6 Study Guide and Deliverables
Text Mining

Readings:

- Online lecture material topics:
  - Latent Semantic Indexing
  - Topic Models
  - Latent Dirichlet Allocation
  - Spark ML Library for NLP

Assignments:

- Term Project due Tuesday, February 27 at 6:00 AM ET

Assessments:

- None

Course Evaluation:

- Course Evaluation opens on Monday, February 26 at 10:00 AM ET and closes on Sunday, March 3 at 11:59 PM ET.
- Please complete the course evaluation. Your feedback is important to MET, as it helps us make improvements to the program and the course for future students.

Live Classroom:

- Tuesday, February 20 from 5:30–7:00 PM ET
- Thursday, February 22, from 5:30–7:00 PM ET
- Live Office (facilitator sessions): to be scheduled each weekend.
- Final Exam Prep session: to be scheduled.

Final Exam Details

The Final Exam is a proctored exam available from **Wednesday, February 28, at 6:00 AM ET to Saturday, March 2, at 11:59 PM ET**. The Computer Science department requires that all final exams be administered using an online proctoring service called Examity, which you will access via your course in Blackboard. In order to take the exam, you are required to have a working webcam and computer that meets Examity's system requirements. A detailed list of those requirements can be found on the How-to Schedule page. Additional information regarding your proctored exam will be forthcoming from the Assessment Administrator. You will be responsible for scheduling your own appointment within the defined exam window.

The Final Exam will be an open book**/open notes** and is accessible only during the final exam period. You can access it from the Assessments section of the course. Your proctor will enter the password to start the exam.

Final Exam duration: **three hours**.

The exam features a combination of multiple-choice, essay, and file-response questions.

Grading Information

Please check the **Study Guide** in the syllabus for Live Classroom dates and specific due dates for assignments and assessments.

Grading Structure and Distribution

The grade for the course is determined by the following:

Activity	Percentages
5 x Homework Assignments	40%
5 x Weekly Quizzes	20%
Term Project and Presentation	10%
Final Exam	30%

Assignments

Homework assignments are focused on applying theory learned in the week's module to a set of data and analyzing that data in PySpark. Weekly homework assignments will focus on implementation of data processing and machine learning algorithms in Apache Spark (PySpark). You will use Google Cloud to run your Spark code on large data sets.

Free of charge usage credits for Google Cloud will be provided through Education accounts.

Due Time: At the end of each module (Please check the Study Guide or the Syllabus for the specific due date).

Where to submit: The "Assignments" section in the left-hand course menu.

Weekly Quizzes

Quizzes will evaluate students' understanding of concepts presented in the corresponding week's module. Students should ensure adequate preparation before starting the quiz. It will not be possible to do well on the quiz without first reviewing the course material in depth and attempting to understand all examples and test yourself questions. It is recommended that you complete the quiz after you feel comfortable with the material and have asked any questions that you may have had.

Due time: at the end of each module (Please check the Study Guide or Syllabus for the specific due date).

Where to complete: The "Assessments" section in the left-hand course menu.

Term Project and Presentation

At the end of this course, you will work on your own Big Data project. You will work on a large data set and analyze and train machine learning algorithms. You will present your project in the form of a 15-minute online presentation. Clear project development guidelines will be provided in the course content in the "Assignment" section.

Final Exam

There will be a proctored Final Exam in this course using a proctoring service called Examity. Detailed instructions regarding your proctored exam will be forthcoming from the Assessment Administrator. You will be responsible for scheduling your own appointment.

Translation Between Letter Grades and Percentages

A (Excellent)	95-100
A- (Excellent; minor improvement needed)	90-94.99
B+ (Very good)	87-89.99
B (Good)	83–86.99
B- (Good; some improvements needed)	80-82.99
C+ (Satisfactory; some significant improvements needed)	77-79.99
C (Satisfactory; significant improvements needed)	73–86.99

C- (Satisfactory; significant improvements required)	70-82.99
D (Many significant improvements required)	65
Unacceptable	0

* The grading scale may be adjusted at the discretion of the course instructor.

Lateness

We recognize that emergencies occur in professional and personal lives. If an emergency occurs that prevents your completion of homework by a deadline, please notify your facilitator or instructor. This must be done in advance of the deadline (unless the emergency makes this impossible, of course). Additional documentation may be requested. Work submitted late without any reason provided will result in a grade deduction: we want to be fair to everyone in this process, including the vast majority of you who sacrifice so much to submit your homework on time in this demanding schedule.