Syllabus

This is a single, concatenated file, suitable for printing or saving as a PDF for offline viewing. Please note that some animations or images may not work.

Description

This module is also available as a concatenated page, suitable for printing or saving as a PDF for offline viewing.

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, 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 basic systems and techniques to store large volumes of data, as well as modern systems for cluster computing based on MapReduce patterns such as Hadoop MapReduce, Apache Spark, and Flink.

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.

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.

Technical Notes

The table of contents expands and contracts (+/- sign) and may conceal some pages. To avoid missing content pages, you are advised to use the next/previous page icons in the top right corner of the learning modules.

This course requires you to access files such as Word documents, PDFs, and/or media files. These files may open in your browser or be downloaded as files, depending on the settings of your browser.

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 Visiting Research Professor at Boston University and Head of the Department of Information systems and network technologies at the Faculty of Computer Science and Engineering—ss. 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 in his tenure, the Faculty has become the largest technical Faculty in Macedonia. Dimitar Trajanov is the leader of the Regional Social Innovation Hub 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 170 journal and conference papers and seven books. He has been involved in more than 70 research and industry projects, of which in more than 40 projects as a project leader.

Original Course Developer: Kia Teymourian, Ph.D.

Computer Science Department

Dr. Kia Teymourian is an Assistant Professor of Computer Science at Boston University's Metropolitan College. Dr. Teymourian holds a PhD from Freie Universität Berlin as well as an MS and BS from Berlin University of Technology (TU-Berlin). His computer science expertise lies in data stream processing and complex event processing, big data programming, semantic technologies, and knowledge representation, as well as web technologies and natural language processing. He has made important contributions to multiple large and international research projects, including several funded by the European Commission, the German Federal Ministry of Education and Research (BMBF), and the DARPA Pliny Project at Rice University. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE), and a member of the Association for Computing Machinery (ACM). At Metropolitan College, Dr. Teymourian teaches data analysis and visualization, as well as software design patterns.

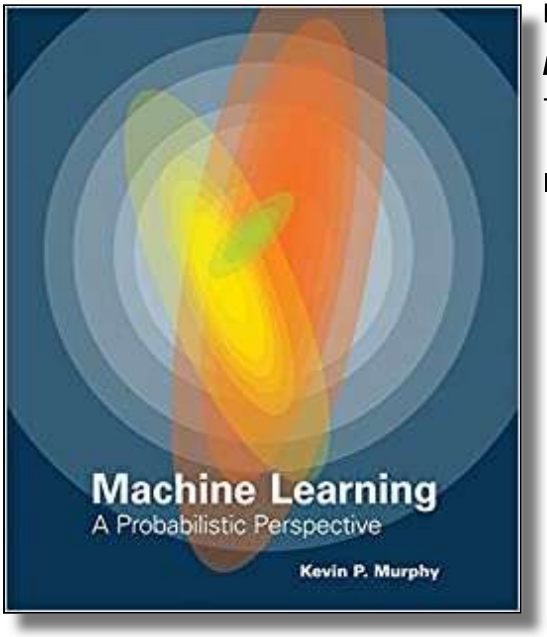
Additional information can be found on Dr. Teymourian's academic website.

Materials

Required Book

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

Recommended Books



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

The MIT Press

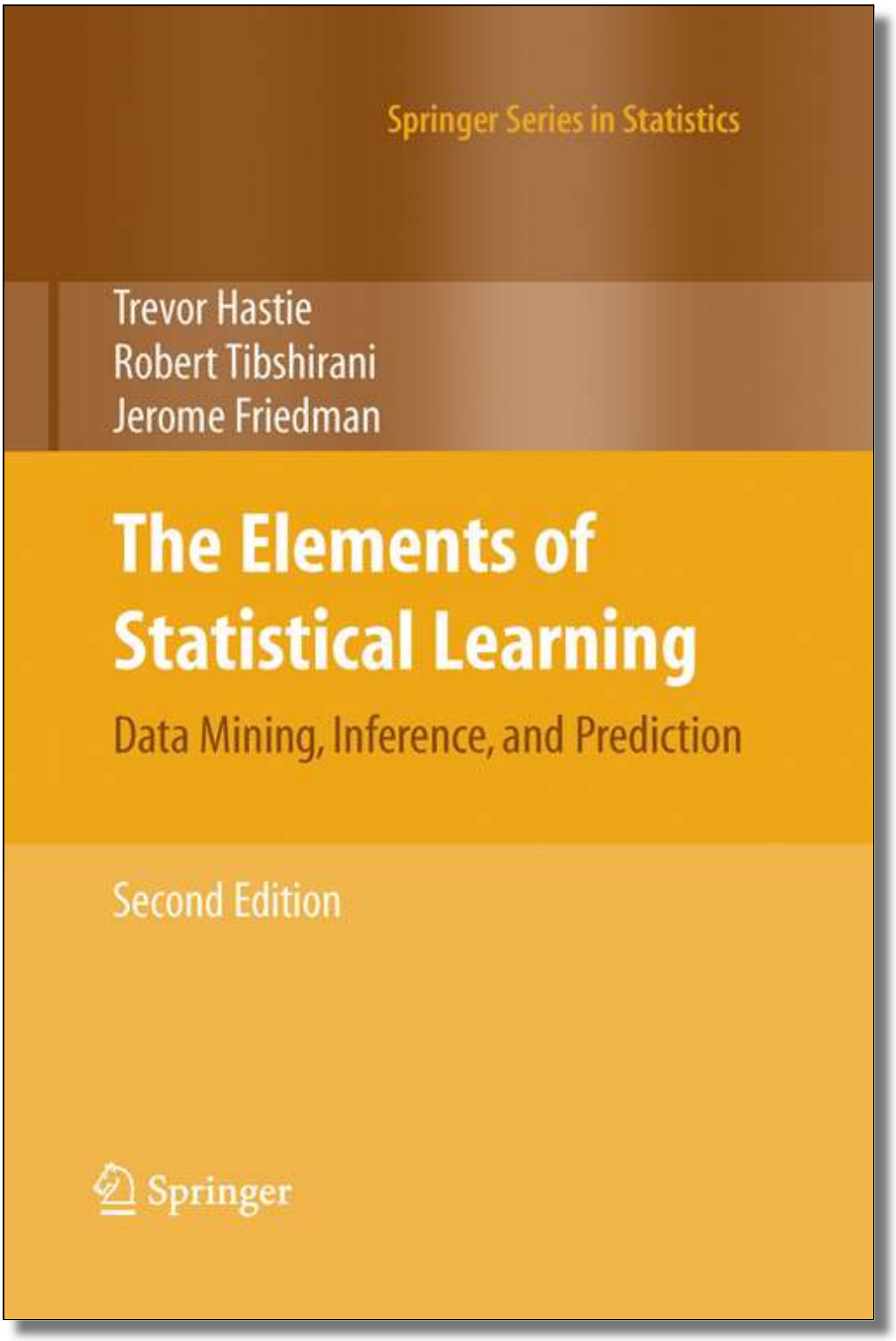
ISBN-13: 978-0262018029

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

Springer-Verlag.

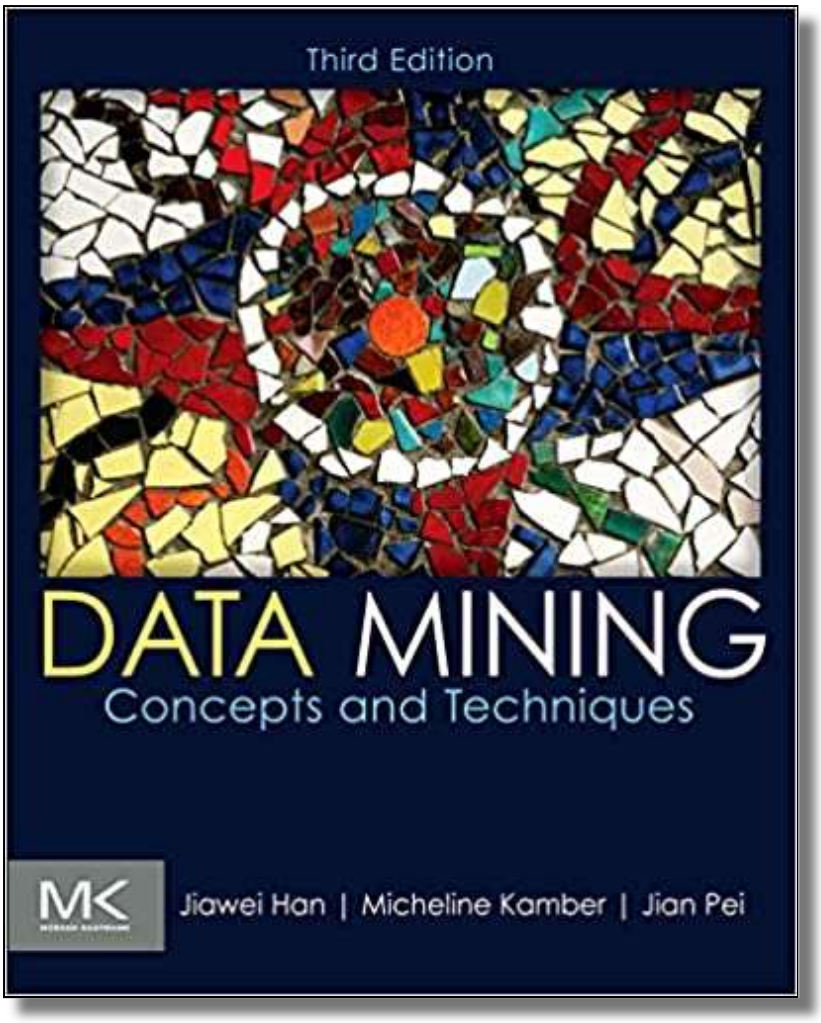
ISBN-13: 978-0-387-84858-7

This book is available for PDF download.



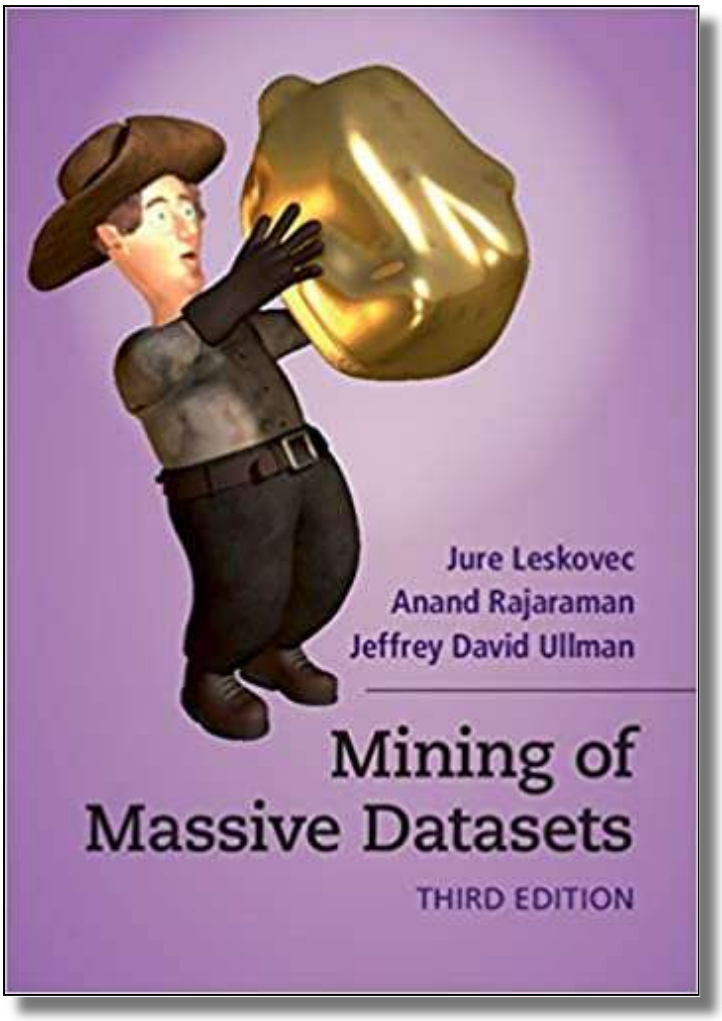
Han, J., Kamber, M., Pei, J. (2009). ***Data mining: Concepts and techniques*** **(3rd ed.).**
Morgan Kaufmann.

ISBN-13: 978-9380931913



Leskovec, J. Rajaraman, A., Ullman, J. (2014). *Mining of massive datasets*.
Cambridge University Press.

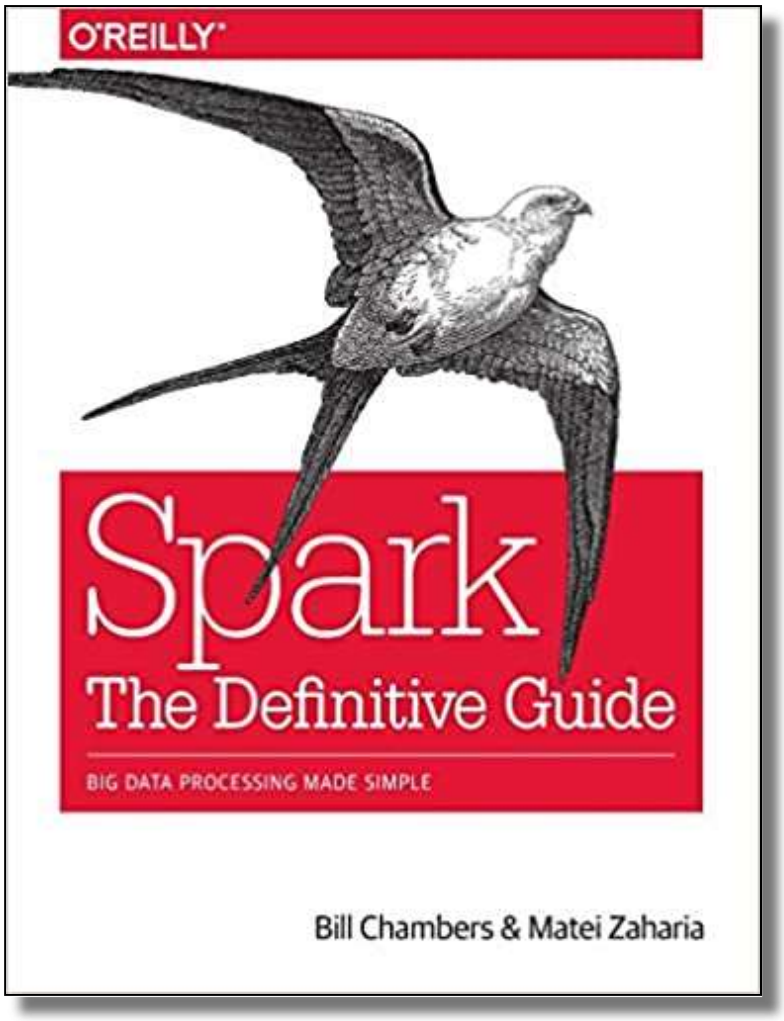
By agreement with the publisher, you can download the book for free from this page.



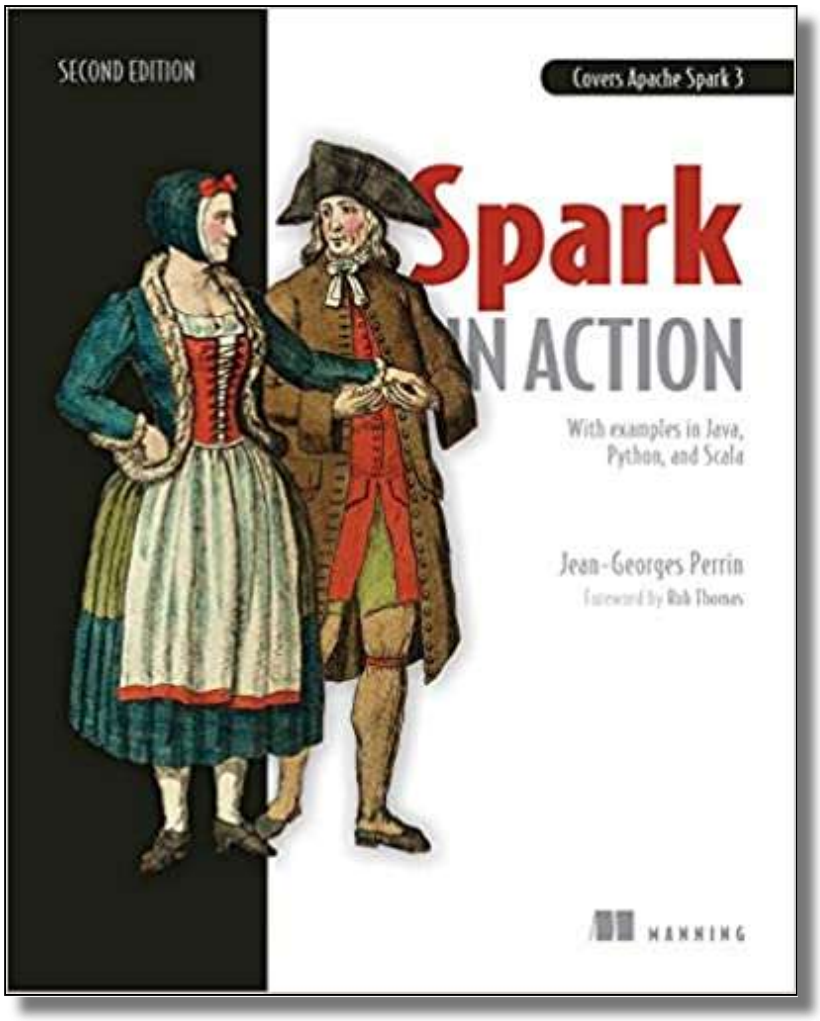
Other Materials and Resources

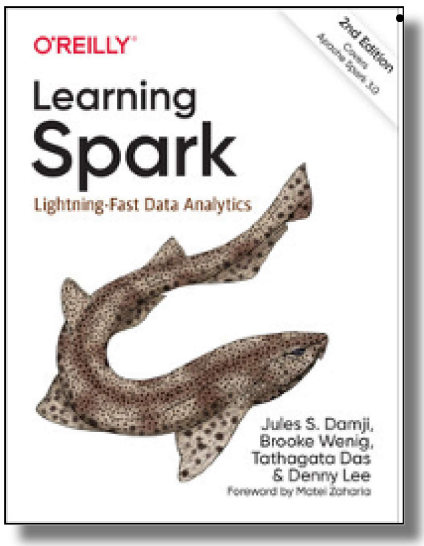
Spark Programming

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

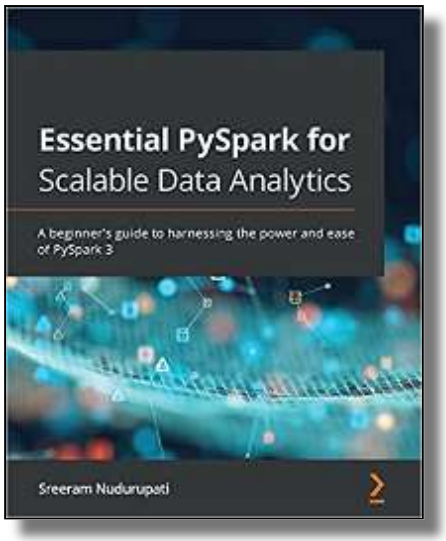


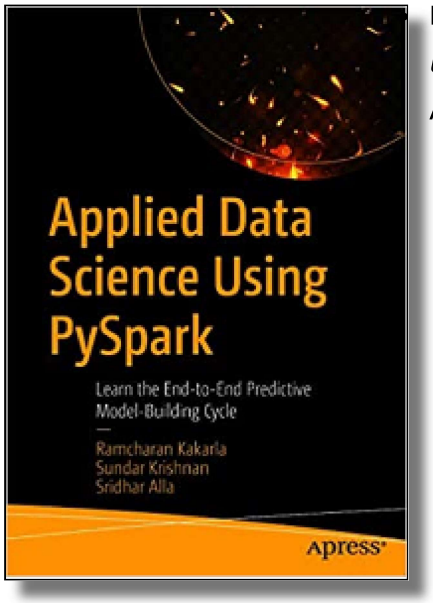
- Perrin, J. (2020). *Spark in action* (2nd ed.). (Covers Apache Spark 3 with examples in Java, Python, and Scala)
  O'Reilly Media Inc.





- 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** for 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 so you can 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) work, 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.

Boston University Library Information

Boston University has created a set of videos to help orient you to the online resources at your disposal. An introduction to the series is below:

> met_ode_library_14_sp1_00_intro video cannot be displayed here

All of the videos in the series are available on the Online Library Resources page, which is also accessible from the Campus Bookmarks section of your Online Campus Dashboard. Please feel free to make use of them.

As Boston University students, you have full access to the BU Library. From any computer, you can gain access to anything at the library that is electronically formatted. You may use the library's content whether you are connected through your online course or not, by confirming your status as a BU community member using your Kerberos password.

Once in the library system, you can use the links under "Resources" and "Collections" to find databases, eJournals, and eBooks, as well as search the library by subject. Some other useful links follow:

> Go to Collections to access eBooks and eJournals directly.
>
> If you have questions about library resources, go to Ask a Librarian to email the library or use the live-chat feature.

To locate course eReserves, go to Reserves.

Please note that you are not to post attachments of the required or other readings in the water cooler or other areas of the course, as it is an infringement on copyright laws and department policy. All students have access to the library system and will need to develop research skills that include how to find articles through library systems and databases.

Free Tutoring Service

Free online tutoring with Smarthinking is available to BU online students for the duration of their courses. The tutors do not rewrite assignments, but instead teach students how to improve their skills in the following areas: writing, math, sciences, business, ESL, and Word/Excel/PowerPoint.

You can log in directly to Smarthinking from Online Campus by using the link in the left-hand navigation menu of your course.

YouTube

Please Note

Smarthinking may be used only for current Boston University online courses and career services. Use of this service for purposes other than current coursework or career services may result in deactivation of your Smarthinking account.

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 Wednesday, May 18 at 6:00 AM ET
Assessments:	• Quiz 1 due Tuesday, May 17 at 6:00 AM ET
Live Classroom:	• Tuesday, May 10 from 5:30–7:00 PM ET • Thursday, May 12 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 + SciPy, Code Optimization, Cluster Configurations
	  ◦ Linear Algebra Computation in Large Scale.
	  ◦ Distributed File Storage Systems

Assignments:	• Assignment 2 due Wednesday, May 25 at at 6:00 AM ET
Assessments:	• Quiz 2 due Tuesday, May 24 at at 6:00 AM ET
Live Classroom:	• Tuesday, May 17 from 5:30–7:00 PM ET • Thursday, May 19 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,
  - Markov Chain Monte Carlo (MCMC)

Assignments:
- Assignment 3 due Wednesday, June 1 at 6:00 AM ET

Assessments:
- Quiz 3 due Tuesday, May 31 at at 6:00 AM ET

Live Classroom:
- Tuesday, May 24 from 5:30–7:00 PM ET
- Thursday, May 26 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
  - Support Vector Machine (SVM) and the kernel trick
  - Outlier Detection
  - Spark ML library

Assignments:
- Assignment 4 due Wednesday, June 8 at at 6:00 AM ET

Assessments:
- Quiz 4 due Tuesday, June 7 at 6:00 AM ET

Live Classroom:	• Tuesday, May 31 from 5:30–7:00 PM ET • Thursday, June 2 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, June 14 at 6:00 AM ET • Assignment 5 due Wednesday, June 15 at at 6:00 AM ET
Assessments:	• Quiz 5 due Tuesday, June 14 at at 6:00 AM ET
Live Classroom:	• Tuesday, June 7 from 5:30–7:00 PM ET • Thursday, June 9 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, June 21 at 6:00 AM ET
Assessments:	• None

Course Evaluation:	Course Evaluation opens on Tuesday, June 14, at 10:00 AM ET and closes on Tuesday, June 21, 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, June 14 from 5:30–7:00 PM ET • Thursday, June 16 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, June 22, at 6:00 AM ET to Saturday, June 25, at 11:59 PM ET**. The Computer Science department requires that all final exams be administered using an online proctoring service called Examity that 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 **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:

Overall Grading Percentages	
5 Homework Assignments	40%
5 Weekly Quizzes	20%
1 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, analyze and train machine learning algorithms. You will present your project in the form of a 15 minutes 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 proctor 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

Lateness

We recognize that emergencies occur in professional and personal lives. If If an emergency occurs that prevents your completion of homework by a deadline, please notify your facilitatorinstructor. 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.