MIDS, UC Berkeley
DATASCI W261
W261 "Machine Learning at Scale"

Course Architect: Dr. James G. Shanahan, 1/1/2015
Revised on April 3, 2017 by Jimi Shanahan

Course Description:
Learn the underlying principles required to develop scalable machine-learning pipelines for structured and unstructured data at the petabyte scale, and gain hands-on experience in Apache Hadoop and Apache Spark and related MapReduce frameworks.

Course Summary:

Until recently, "big data" was very much the purview of database management and summary statistics systems such as Hadoop (HDFS and MapReduce) and was largely underleveraged by machine learning. These systems, though useful, suffered from their limited utility. This course builds on and goes beyond this collect-and-analyze phase of big data by focusing on how machine learning algorithms can be rewritten and in some cases extended to scale to work on petabytes of data, both structured and unstructured, to generate sophisticated models that can be used for real-time predictions. Predictive modeling at this scale can lead to huge boosts in performance (typically in the order of 10–20%) over small-scale models running on stand-alone computers that require one to significantly down-sample and, necessarily, to simplify big data. Concretely, this course focuses on how the MapReduce design pattern from parallel computing can be extended and more faithfully leveraged to tackle the somewhat "embarrassingly parallel" task of machine learning (a lot of machine learning algorithms fit this mold). In this course, this is accomplished via the Apache Spark project and its many related subprojects. Apache Spark is an open-source cluster-computing framework. It has emerged as the next-generation big data processing engine, overtaking Hadoop MapReduce, which helped ignite the big data revolution. Spark maintains MapReduce's linear scalability and fault tolerance but extends it in a few important ways: it is much faster (100 times faster for certain applications); much easier to program in, due to its rich APIs in Python, Java, and Scala (and R) and its core data abstraction, the distributed data frame; and it goes far beyond batch applications to support a variety of compute-intensive tasks, including interactive queries, streaming, machine learning, and graph processing.

This course will provide an accessible introduction to MapReduce frameworks and to Spark and its potential to revolutionize academic and commercial data science practices through scale. Conceptually,

the course is divided into two parts. The first part will cover fundamental concepts of MapReduce parallel computing via Hadoop, MRJob, and Spark, while diving deep into Spark core, data frames, the Spark shell, Spark streaming, Spark SQL, MLlib, and more. The second part will focus on hands-on algorithmic design and development in parallel computing environments such as Spark; developing algorithms from scratch, such as decision-tree learning; graph-processing algorithms such as PageRank and shortest path; gradient descent algorithms such as support vectors machines; and matrix factorization. Industrial applications and deployments of MapReduce parallel compute frameworks from various fields, including advertising, finance, healthcare, and search engines, will also be presented. Examples and exercises will be made available in Python notebooks (Hadoop streaming, MRJob, and PySpark).

Prerequisites: Introductory machine learning course or equivalent. Intermediate programming capabilities in an object-oriented language (such as Python).

Course Assignments and Exams:

There will be homework associated with each week of class in this course. These assignments will be completed individually. The midterm exam will cover all the major topics in the course to that point. Finally, there is a group project that you will complete with a small team, using one of the course data sets.

% of Final Grade	Task
40%	Homework+Projects (iterate/agile): Projects will have phases (all structured)
20%	Midterm Exam (Week 8 of the semester)
20%	Class participation (fact-to-face and in the online forums; please answer each other's questions; collaborate; communicate)
20%	End term exam (Week 14 of the Semester)

Homework Schedule and policy

- Homework submissions are due on Tuesday morning by 8AM, West Coast Time (this is a strict deadline as we have a tough schedule to keep up with each week during semester.
- Each day you are late you will incur a **33.33% penalty.**
- Grade each HW and give written feedback to students by Sunday night of each week. This is a great point of calibration. TAs should make instructors aware of student performances and and areas where students did not do well.
- Publish HW results on Monday, along with master solution
- E.g., (with all times in US West Coast Time)
    - HW1 is due on the Tuesday of Week 2 at 8AM (Sharp)
    - HW1 will be graded by Sunday night of Week 2
    - Grades and master solution for HW1 will be published (on ISVC) by Sunday midnight of Week 2.

Homework Location

- Github Repo:
- https://classroom.github.com/assignment-invitations/562761d1f7fde3700357401ff5ad25fa

Office Hours

Office hours are an on an on-demand basis. This need will be met largely by extra time after live sessions If required we can do additional office on **Sundays, at 2PM West Coast Time.**

NOTE: You can only attend if you submit a question(s) 24 hours before the start of office hours to the online forum (SLACK). Use the STAR methodology to present you Questions/Answers for both Knowledge based questions and for problem solving questions

Slack Channel

Instructors Channel: https://ucbischool.slack.com/messages/w261-instructors

Students Channel: https://ucbischool.slack.com/messages/w261-summer-2017

- Join Slack Channel for questions/discussion
- JimiUCB@ucbischool.slack.com, Mike.Tamir, KyleHamilton
  - #2017-0109_w261
  - https://ucbischool.slack.com/messages/2017-0109_w261/

Other important Links

- Calendar and Google Drive:
  - https://docs.google.com/spreadsheets/d/1RRmGU5aymKWbFjxvK_ElQJnfTaB-G2U3ByNv2b0wWO0/edit#gid=1565460380
- Slides and other folders
  - https://drive.google.com/open?id=0B0b2FfTVAWAMTDZOWnN3VDhUOWc

Course Software:

Python and iPython Notebooks will be the common link between all the MapReduce frameworks we will be using to teach with and to complete projects. Having said that, the core focus of this class is large-scale machine learning. As such, we will be using frameworks that allow the handling of petabytes

of data (both structured and unstructured) from the following perspectives: storage, network transfer, exploratory data analysis, modeling, and experimentation. We will primarily focus on the following MapReduce frameworks:

- Hadoop streaming [Week 1]
- MRJob [Week 3]
- Spark and some of its specialty libraries (SparkSQL, MLlib, GraphX) [Week 10]
- AltaScale [Week 4 introduction]

Textbooks and Reader:

This course will use a combination of textbooks and some online readings. You can buy the paper or electronic copies of the textbooks. All students are expected to do the readings before watching a asynchronous lecture and before all synchronous (also known as live) sessions. In many cases, the lecture materials and the weekly discussion sessions will not make sense if you do not have knowledge of the examples and issues from the readings.

Recommended Textbooks:

Karau, Holden, Konwinski, Andy, Wendell, Patrick, & Zaharia, Matei. (2015). *Learning Spark: Lightning-fast big data analysis*. Sebastopol, CA: O'Reilly Publishers.

Lin, Jimmy, & Dyer, Chris. (2010). *Data-intensive text processing with MapReduce*. San Rafael, CA: Morgan & Claypool Publishers. (Free online)

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Stanford, CA: Springer Science+Business Media. (Free online)

Ryza, Sandy, Laserson, Uri, Owen, Sean, & Wills, Josh. (2015). *Advanced analytics with Spark: Patterns for learning from data at scale*. Sebastopol, CA: O'Reilly Publishers.

Leskovec Jure, Rajaraman Anand, Ullman jeff, (2014). *Mining of Massive Datasets*, Cambridge University Press. Book available online at http://www.mmds.org/

Online references: Other references such as research pages, webpages and examples are included in the lectures directly.

Course Logistics:

Every week you will cover lecture 1 by yourself asynchronously. The goal of lecture 2 (also known as the live session) is for us to meet and discuss questions that you may have, review homework and also to reinforce the topics in lecture 1 with additional examples or background material. If time permits we

may occasionally build on the lecture 1 topics and introduce new and related topics.

Syllabus

Unit 1	Introduction and Motivation for Machine Learning at Scale
Lecture 1	**Introduction and Motivation for Machine Learning at Scale**
References	Data Science ● Selections from, Doing data science, Schutt and O'Neil; (Chapters 1 and 2) ● Youtube video: The human face of big data: Rick Smolan at TEDxMidwest (https://www.youtube.com/watch?v=h8FLkRK-wF4) Excellent articles with code in matlab for bias-variance analysis of squared error loss ● https://theclevermachine.wordpress.com/2013/04/21/model-selection-underfitting-overfitting-and-the-bias-variance-tradeoff/ ● http://insidebigdata.com/2014/10/22/ask-data-scientist-bias-vs-variance-tradeoff/ Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Stanford, CA: Springer Science+Business Media. (Chapter 7) Also other referenced materials (online links are embedded in slides); this is the case for all lectures
Topics	Big data Sources of big data Data Scientist Large scale machine learning Data modeling pipeline Bias-variance tradeoff Back of envelope calculations A MapReduce framework using Unix Command line
Lecture 2	**A command line map-reduce environment (Lecture 2 corresponds to the weekly synchronous time period; we may cover some or all of the topics listed here depending on the available time and needs of the students)**
References	Introduction to information retrieval; Manning, Raghavan, & Schutz, Chapter 13 (Naïve Bayes) https://en.wikipedia.org/wiki/Reservoir_sampling
Topics	Self-Introductions (Each student: please prepare 100-200 word bio of yourself to paste into the chat pod in Adobe Connect) Bernoulli Naive Bayes Multinomial Naive Bayes Coding up a distributed Naive Bayes using Python (and Unix command-line utilities)

Unit 2	Parallel Computing, MapReduce, and Hadoop (Data Storage and Algorithms)
Lecture 1	**Parallel Computing, MapReduce, and Hadoop (Data Storage and Algorithms)**
References	● Lin, Jimmy, & Dyer, Chris. (2010). Data-intensive text processing with MapReduce. San Rafael, CA: Morgan & Claypool Publishers. (Chapter 2) ● http://en.wikipedia.org/wiki/Barrier_(computer_science ● http://en.wikipedia.org/wiki/Embarrassingly_parallel ● http://gerardnico.com/wiki/data_storage/shared_disk_nothing ● http://en.wikipedia.org/wiki/Shared_nothing_architecture ● http://en.wikipedia.org/wiki/Message_Passing_Interface ● http://en.wikipedia.org/wiki/Map_(parallel_pattern) ● http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/ ● http://hadoop.apache.org/core/docs/current/ ● See also other referenced material
Topics	● Motivation for parallel computing ● Parallel computing ● Parallel Computing (PC) Definition and Communication Synchronization Types of PC Tasks ● Architectures for Parallel Computation ● Developer Frameworks for Parallel Computation ● Hadoop Background and History ● Hadoop Distributed File System (HDFS) ● MapReduce: Functional Programming ● Hadoop: MapReduce ● Animated Examples
Lecture 2	**Installing and Programming Hadoop**
References	Hadoop online documentation
Topics	● MapReduce: Runtime environment ● Hadoop 2.0 ● Install Hadoop; run locally; run in the cloud? ● Run wordcount notebook and explain each step ● 10 most popular words ● Examples of problem solving in Hadoop ● Naive Bayes in Hadoop

Unit 3	MapReduce Algorithm Design and Design Patterns
Lecture 1	**MapReduce Algorithm Design and Design Patterns**
References	• Lin, Jimmy, & Dyer, Chris. (2010). Data-intensive text processing with MapReduce. San Rafael, CA: Morgan & Claypool Publishers. (Chapter 3). • https://en.wikipedia.org/wiki/Priority_queue • **http://en.wikipedia.org/wiki/Merge_algorithm** See also other referenced material
Topics	• Background • RAM vs. disk vs. bandwidth • Priority queues and merge sort • The internals of the Hadoop shuffle • Local aggregation • Combiners and in-mapper combining • Algorithmic correctness with local aggregation • How many Mappers and Reducers? • Pairs and stripes • Computing relative frequencies
Lecture 2	**Hadoop at Scale and Installing MRJob**
References	• Introduction to information retrieval; Manning, Raghavan, & Schutz, Chapter 13 • Lin, Jimmy, & Dyer, Chris. (2010). Data-intensive text processing with MapReduce. San Rafael, CA: Morgan & Claypool Publishers. (Section 3.4). • https://en.wikipedia.org/wiki/Apriori_algorithm • MMDS Book on Data mining, o Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeff Ullman, Book available online at http://www.mmds.org/ o https://www.dropbox.com/s/yoapjbpoiwmfssy/MMDS-Book-Chapter6.pdf?dl=0
Topics	• Secondary sorting • Coding and debugging Hadoop • A-Priori algorithm for mining frequent itemsets (single brutefore map-reduce job for itemsets 1… n, and for association rule extraction as opposed to a multistep pipeline with that has iterative part) • Installing MRJob

Unit 4	Unsupervised Learning at Scale in MapReduce: Clustering, K-Means, Expectation Maximization
Lecture 1	**MRJob, Unsupervised Learning at Scale: Clustering and K-Means**

References	• Introduction to information retrieval; Manning, Raghavan, & Schutz, Chapters 16 and 17
	• MRJob online documentation (https://pythonhosted.org/mrjob/)
	• MrJob examples
	  o http://pythonhosted.org/mrjob/guides/writing-mrjobs.html%23writing-basics
	  o https://github.com/uchicago-cs/cmsc12300/tree/master/examples/data_analysis/bin
	• Microsoft customer visitor log file data
	  o http://archive.ics.uci.edu/ml/machine-learning-databases/anonymous/
	• Google N-Gram data
	  o **http://storage.googleapis.com/books/ngrams/books/datasetsv2.html**
	• **Yelp review challenge:**
	  o **http://www.yelp.com/dataset_challenge/**
	  o **https://github.com/jblomo/pycon-mrjob/blob/master/code/unique_review.py**
	• See also other referenced material

	• MRJob Installation • MRJob fundamentals and concepts • Writing MRJob code • Log-file processing • Serializable, JSON, and other MRJob information • Clustering overview • K-means algorithm
Lecture 2	**Model-Based Clustering: Expectation Maximization**

References	• Introduction to information retrieval; Manning, Raghavan, & Schutz, Chapter 16
	• MMDS Book on Data mining,
	  o Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeff Ullman, Book available online at http://www.mmds.org/
	  o https://www.dropbox.com/s/yoapjbpoiwmfssy/MMDS-Book-Chapter6.pdf?dl=0

Topics	• Distributed k-means in MRJob • Distributed k-means in MRJob (code) • Data and distributions • Model-based clustering • EM algorithm for a mixture of Bernoulli distributions • Distributed EM clustering algorithm in MRJob • *Initialization (canopy clustering) if time permits*

Unit 5	Big Data Pipelines

Lecture 1	**Big Data Pipelines**
References	• Lin, Jimmy, & Dyer, Chris. (2010). Data-intensive text processing with MapReduce. San Rafael, CA: Morgan & Claypool Publishers. (Chapter 3); • Shanahan, James G., & Kurra, Goutham. (2011). Digital advertising: An information scientist's perspective. In Massimo Melucci & Ricardo Baeza-Yates (Eds.), Advanced topics in information retrieval, pp. 209–238. Berlin and Heidelberg: Springer-Verlag • **http://en.wikipedia.org/wiki/Data_warehouse** • https://en.wikipedia.org/wiki/Join_(SQL) See also other referenced material
Topics	• Mobile advertising example • Data warehousing, OLTP/OLAP • Database relational joins • Databases Operations in MapReduce • Relational Joins • Reduce Side Join • Reduce Side Join Example • Memory Backed Map Side Join • Map Side Join Merge Join
Lecture 2	**Exploratory Data Analysis and Evaluation and Distributed Information Retrievlal**
References	• Lin, Jimmy, & Dyer, Chris. (2010). Data-intensive text processing with MapReduce. San Rafael, CA: Morgan & Claypool Publishers. (Chapter 4); • https://en.wikipedia.org/wiki/Reservoir_sampling • https://en.wikipedia.org/wiki/Column-oriented_DBMS

Topics	• Counters in MrJob/Hadoop
	• Online advertising
	  o Demand-side platform challenges
	  o Case study (DSP competition)
	• Column oriented databases
	• Distributed information retrieval? Why a new database structure?
	• Summary statistics
	• Sampling

Unit 6	Supervised Machine Learning in MapReduce (Part 1: Linear Regression)
Lecture 1	**Distributed Supervised Machine Learning (Distributed Linear Regression via Two Approaches)**
References	• Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Stanford, CA: Springer Science+Business Media. (Chapter 3). • https://en.wikipedia.org/wiki/Machine_learning • Lin, Jimmy, & Dyer, Chris. (2010). Data-intensive text processing with MapReduce. San Rafael, CA: Morgan & Claypool Publishers. (Section 3.4). • https://en.wikipedia.org/wiki/Mathematical_optimization • https://en.wikipedia.org/wiki/Convex_optimization • http://en.wikipedia.org/wiki/Gradient • MapReduce Study (http://www.cs.stanford.edu/people/ang/papers/nips06-mapreducemulticore.pdf) See also other referenced material
Topics	• Supervised machine learning (parametric vs. nonparametric) • Applications of matrices at scale • Distributed matrix and vector multiplication • Optimization theory • Gradient descent • Convex optimization • Linear regression • Distributed closed-form linear regression • Complexity, scaling out, and case study • Distributed linear regression via gradient descent
Lecture 2	**Regression Diagnostics, Experiments, and Extensions of Linear Regression**
References	Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Stanford, CA: Springer Science+Business Media. (Chapter 3).
Topics	• Ridge and Lasso regression (part 1) • Nonparametric approaches to machine learning at scale • Evaluation of models (internal diagnostics and metrics) • A-B testing

Unit 7	Introduction to Graph Algorithms at Scale: Single Shortest Path Algorithm
Lecture 1	**Graph-Based Algorithms**
References	• Lin, Jimmy, & Dyer, Chris. (2010). *Data-intensive text processing with MapReduce*. San Rafael, CA: Morgan & Claypool Publishers. (Chapter 5). • **Dijkstra's_algorithm** o http://www.cs.cornell.edu/courses/cs312/2002sp/lectures/lec20/lec20.htm o http://en.wikipedia.org/wiki/Dijkstra's_algorithm • See also other referenced material
Topics	• Networks Introduction and Motivation • Applications of Graphs • Graph Definitions • Shortest Path Introduction • Single Source Shortest Path (SSSP) Unweighted Graphs • BFS for Unweighted Graphs Directed and Animations • SSSP for Weighted Graphs BFS • Dijkstras Algorithm
Lecture 2	**Implementing graph algorithms**
References	https://en.wikipedia.org/wiki/Connected_component_(graph_theory)
Topics	• SSSP • Connected-component • Implementing Connected components • Q&A and planning for mid term

Unit 8	Supervised Machine Learning in MapReduce (Part 1: Linear Regression)
Lecture 1	**Mid-Term Exam**
References	Weeks 1–7 reading material
Topics	A three-hour open-book exam will be administered based on the course material to date
Lecture 2	**Mid-Term Exam**
References	----
Topics	----

Unit 9	Large Scale Graph Processing: Random Walks, PageRank, Personalized PageRank
Lecture 1	**Large Scale Graph Processing: Random Walks, PageRank, Personalized PageRank**
References	Lin, Jimmy, & Dyer, Chris. (2010). *Data-intensive text processing with MapReduce*. San Rafael, CA: Morgan & Claypool Publishers. (Chapter 5). Overview article on PageRank ● Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). The PageRank citation ranking: Bringing order to the Web ● http://www.dam.brown.edu/people/mchb/la/GooglePageRank.pdf PageRank in Python/MrJob ● https://github.com/Yelp/mrjob/blob/master/mrjob/examples/mr_page_rank.py ● http://www.cs205.org/resources/05-MapReduceDesignPatterns.pdf ● The IR Book Chapter 21, http://nlp.stanford.edu/IR-book/pdf/21link.pdf
Topics	● Web Search Introduction ● WebGraph History ● Random Walk ● Random Walk as a Markov Chain ● PageRank ● PCA and Power Iteration ● Distributed PageRank ● Distributed PageRank Example ● PageRank and Dangling Nodes ● Topic-Specific PageRankcent
Lecture 2	**Implementation and applications of Distributed Pagerank and install Spark**
References	Karau, Holden, Konwinski, Andy, Wendell, Patrick, & Zaharia, Matei. (2015). *Learning Spark: Lightning-fast big data analysis*. Sebastopol, CA: O'Reilly Publishers. (Chapters 1—2)
Topics	● Implementations of Distributed Pagerank and its derivatives ● HITS Algorithm ● TextRank ● Applications of pagerank (e.g., Graph analysis of movies) ● Project planning ● Install Spark ● Final project planning

Unit 10	Apache Spark: From Basics to Advanced
Lecture 1	**Introduction to Apache Spark**
References	Karau, Holden, Konwinski, Andy, Wendell, Patrick, & Zaharia, Matei. (2015). Learning Spark: Lightning-fast big data analysis. Sebastopol, CA: O'Reilly Publishers. (Chapters 1–3) See also other referenced material
Topics	● Functional Progamming Review and lazy evaluation ● Background on Spark ● Spark basics ● Programming with Base RDDs ● Transformation and actions ● Animated Example Data Flow ● Pair RDDs ● Basic WordCount in Spark
Lecture 2	**Advanced Programming in Spark**
References	Karau, Holden, Konwinski, Andy, Wendell, Patrick, & Zaharia, Matei. (2015). Learning Spark: Lightning-fast big data analysis. Sebastopol, CA: O'Reilly Publishers. (Chapter 4
Topics	● Accumulators and broad cast variables ● Joins ● Linear regression in Spark ● Spark at scale (EC2 cluster) ● Scala, tools for EDA, SparkSQL ● Join, statistics in Spark, streaming, Naive Bayes ● Project planning

Unit 11	Distributed Supervised Machine Learning Part 2 (in Spark)
Lecture 1	**Distributed Supervised Machine Learning Part 2 (in Spark)**

References	• Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Stanford, CA: Springer Science+Business Media. (Section 4.4).
	• Izenman, Alan J. (2013). Modern multivariate statistical techniques: Regression, classification, and manifold learning. Springer Texts in Statistics. Stanford, CA: Springer Science+Business Media. (Chapter 11)
	• Loss functions
	  o http://www.ics.uci.edu/~dramanan/teaching/ics273a_winter08/lectures/lecture14.pdf
	• **Distributed Perceptron**
	  o **http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/36266.pdf**
	  o **http://cui.unige.ch/~gesmundo/papers/gesmundo-eacl2012.pdf**
	•

Topics	• Loss Functions and General Framework for Gradient Descent • Logistic Regression at Scale • Perceptrons • Distributed Perceptron and case study • Support vector machines (SVM) • SGD Based SVMs at Scale
Lecture 2	**Regression Diagnostics, Experiments, and Extensions of Linear Regression**
References	• Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Stanford, CA: Springer Science+Business Media. (Section 4.4). • Izenman, Alan J. (2013). Modern multivariate statistical techniques: Regression, classification, and manifold learning. Springer Texts in Statistics. Stanford, CA: Springer Science+Business Media. (Chapter 11)
Topics	• Derive dual form of SVMs • SMO algorithm • Distributed SMO • Kernel-AdaTron • QP-based algorithms • Nonparametric approaches to machine learning at scale • Ridge and Lasso logistic regression • Project work

Unit 12	Non-Gradient-Based Machine Learning Algorithms in Spark and Ensembles (Special Focus on Decision Trees)
Lecture 1	**Non-Gradient-Based Machine Learning Algorithms in Spark (Special Focus on Decision Trees)**
References	• Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Stanford, CA: Springer Science+Business Media. (section 9.2) • https://en.wikipedia.org/wiki/Decision_tree_learning • Planet: Massively Parallel Learning of Tree Ensembles o http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36296.pdf • Ryza, Sandy, Laserson, Uri, Owen, Sean, & Wills, Josh. (2015). Advanced analytics with Spark: Patterns for learning from data at scale. Sebastopol, CA: O'Reilly Publishers. (Chapter 4)
Topics	• Introduction to Decision Trees • Decision Trees Modes Medians Means and Machine Learning • Decision Tree Prediction • Decision Trees Learning a Regression Tree Over Real Valued Variables • Decision Trees Learning a Classification Tree • Decision Trees Understanding Input Variables • Decision Trees Representation • Learning a Decision Tree • Distributed Decision Tree Learning • Planet Overview
Lecture 2	**Ensembles and Machine learning Case Studies in Spark**
References	• Planet: Massively Parallel Learning of Tree Ensembles o http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36296.pdf • Ryza, Sandy, Laserson, Uri, Owen, Sean, & Wills, Josh. (2015). Advanced analytics with Spark: Patterns for learning from data at scale. Sebastopol, CA: O'Reilly Publishers. (chapter 4)
Topics	• Decisions trees at scale • Ensembles of decision trees • MLlib introduction • Data frames in Spark • Case study: Airline delay prediction • Other case studies (advanced Spark book) • Project work

Unit 13	Predicting Links and Attributes in Social Networks (OPTIONAL)
Lecture 1	**Predicting Links and Attributes in Social Networks**
References	• Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 635-644. DOI=10.1145/1935826.1935914 o http://cs.stanford.edu/people/jure/pubs/linkpred-wsdm11.pdf • Label propagation in graphs o S. A. Macskassy and F. Provost. A simple relational classifier. In Proceedings of the Multi-Relational Data Mining Workshop (MRDM) at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003. o *Joint Inference of Multiple Label Types in Large Networks*, pdf ppt, by D. Chakrabarti, S. Funiak, J. Chang, and S. A. Macskassy, in ICML 2014. • TextRank: Bringing Order into Texts, Rada Mihalcea and Paul Tarau, EMNLP Conference 2004 o http://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
Topics	• A Supervised Random Walk • Supervised Random Walk Case Studies • Label Propagation Introduction • Weighted Vote Relational Neighbor Classifier • Facebook Case Study Inferring Missing Node Attributes • TextRank • Apache Big Data Graph Processing Frameworks • Friends of Friends
Lecture 2	**Regression Diagnostics, Experiments, and Extensions of Linear Regression**
References	Ryza, Sandy, Laserson, Uri, Owen, Sean, & Wills, Josh. *Advanced analytics with Spark: Patterns for learning from data at scale*. Sebastopol, CA: O'Reilly Publishers. (Chapter 7)
Topics	• Co-occurrence networks and graphX • Project work

Unit 14	Recommender Systems, PCA, SVD, Matrix Factorization and MLLib
Lecture 1	**Alternating Lease Squares and Various Optimization in Spark/Working with Spark MLLib**
References	Ryza, Sandy, Laserson, Uri, Owen, Sean, & Wills, Josh. *Advanced analytics with Spark: Patterns for learning from data at scale*. Sebastopol, CA: O'Reilly Publishers. (Chapter 3)
Topics	● ALS Introduction ● Single Node ALS Example ● ALS Parallelization ● Distributed ALS Closed Form ● How to Implement ALS in Spark ● Distributed ALS Gradient Descent ● Pregel GraphX ● ALS in Code ● MLlib ● Course Wrap Up
Lecture 2	SVD at Scale via ALS and Case Study
References	Ryza, Sandy, Laserson, Uri, Owen, Sean, & Wills, Josh. *Advanced analytics with Spark: Patterns for learning from data at scale*. Sebastopol, CA: O'Reilly Publishers. (Chapter 6)
Topics	● Recommender systems: Overview ● Item-to-item algorithm ● Project work ● Item-to-item algorithm in Spark and MRJob o MRJob (http://aimotion.blogspot.com/2012/08/introduction-to-recommendations-with.html) ● PCA and Eigenfaces ● Matrix factorization ● ALS vs. SVD ● MovieLens uses all data (not just the people's star ratings but the missing ones also)

Unit 15	Deep Learning (OPTIONAL)
Lecture 1	**Presentations and Discussion**
References	---- Deep Learning references are provided within the slides
Topics	• Deep Learning lecture

Unit 15: Presentation and Discussion of Final Project

Lecture 1: Presentations and Discussion

Lecture 2: Presentations and Discussion