Mining of Massive Datasets
CSE6017

Dr. Ramesh Ragala
VIT Chennai

Syllabus Introduction

- **Course Objective**:
  - To provide comprehensive knowledge on developing and applying machine learning algorithms for massive real-world datasets in distributed frameworks
- **Expected Outcomes**:
  - 1. Identify right machine learning / mining algorithm for handling massive data
  - 2. Implement machine learning algorithms in distributed frameworks such as MapReduce and Spark
  - 3. Use deep learning and extreme learning to solve real-life problems having multifarious complexities
  - 4. Use big data analytics tools such as Spark, Mahout and H_2O in solving problems based on Machine learning

Syllabus Introduction – UNIT -I

- MapReduce Based Machine Learning on
  - K-Means, k-means++
  - PLANET
  - Parallel SVM
  - Association Rule Mining in MapReduce
  - Inverted Index
  - Page Ranking
  - Expectation Maximization
  - Bayesian Networks

Syllabus Introduction – UNIT -II

- Classification & Regression models with Spark & Mahout
  - Linear support vector machines
  - Naive Bayes model
  - Decision Trees
  - Least square regression
  - Decision trees for regression

Syllabus Introduction – UNIT -III

- Clustering in Spark and Mahout
  - Hierarchical Clustering in a Euclidean and Non-Euclidean Space
  - The Algorithm of Bradley, Fayyad, and Reina
  - A variant of K-means algorithm
  - Processing Data in BFR Algorithm
  - CURE algorithm
  - Clustering models with Spark
  - Spectral clustering using Mahout

Syllabus Introduction – UNIT -IV

- Mining Social-Network Graphs
  - Clustering of Social-Network Graphs
  - Direct Discovery of Communities
  - Partitioning of Graphs
  - Finding Overlapping Communities
  - Counting Triangles using MapReduce
  - Neighborhood Properties of Graphs

Syllabus Introduction – UNIT -V

- Semi-Supervised Learning
  - Introduction to Semi-Supervised Learning
  - Semi-Supervised Clustering
  - Transductive Support Vector Machines

Syllabus Introduction – UNIT -VI

- Deep Learning
  - Introduction to Deep Learning
  - Deep Neural Networks
  - Deep Belief Networks
  - Auto Encoders
  - Recurrent Networks

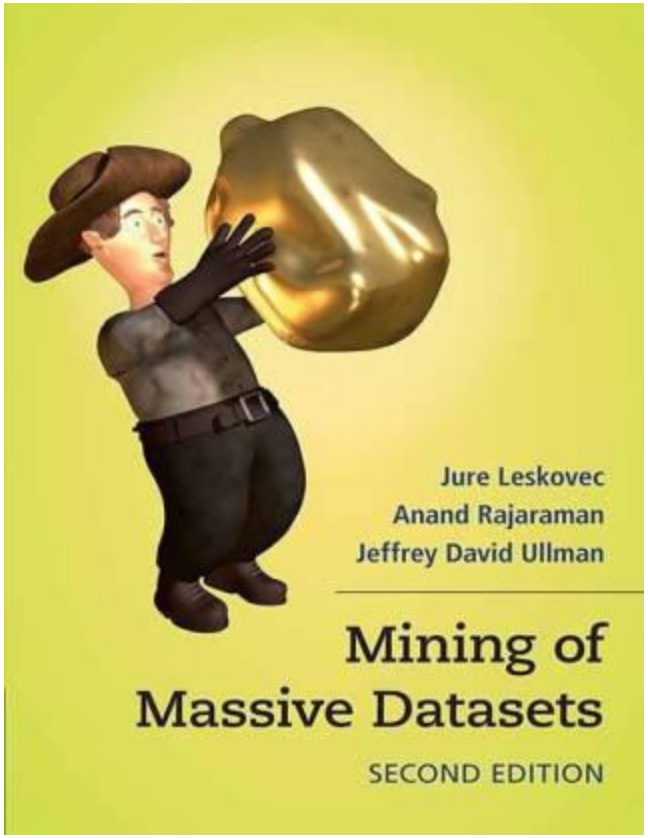
Syllabus Introduction – UNIT -VII

- Extreme Learning Machine
  - Introduction to Extreme Learning Machines (ELM)
  - ELM auto encoder
  - Extreme Support Vector Regression

RECENT TRENDS

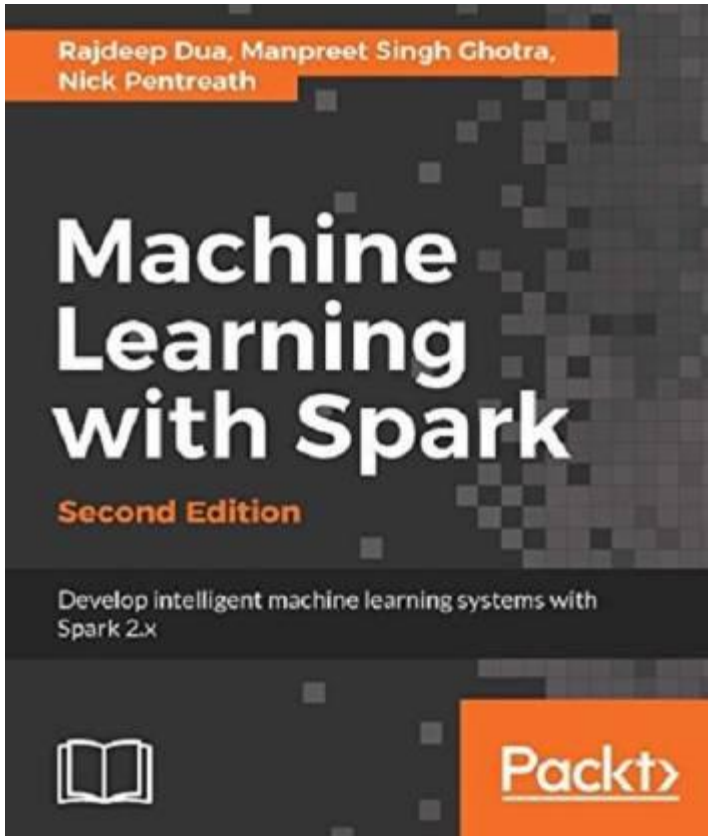
Reference Books

- Jure Leskovec, AnandRajaraman, Jeff Ullman, “Mining of Massive Datasets”, Standford Press,2011



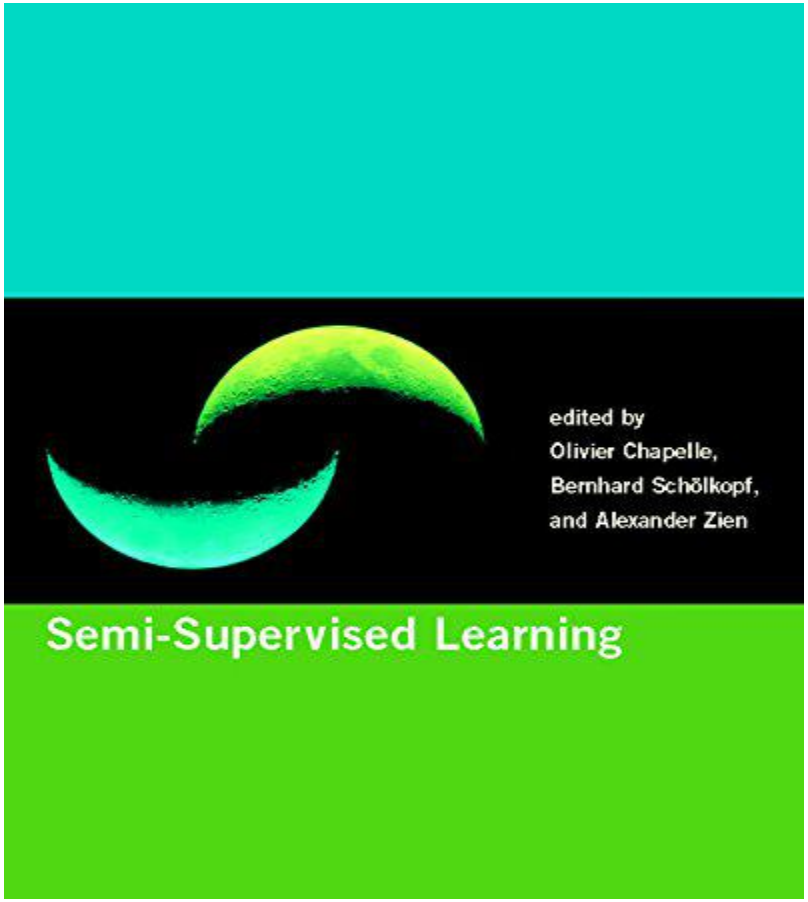
Reference Books

- Nick Pentreath, “Machine Learning with Spark”, Packt Publishing, 2015



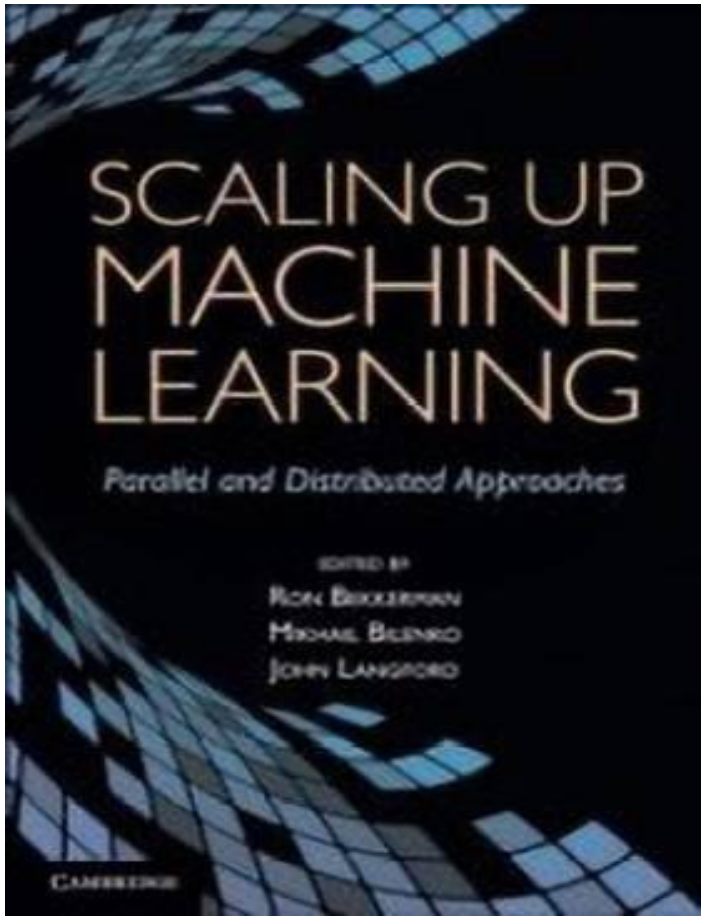
Reference Books

- Olivier Chapelle, Bernhard Scholkopf, Alexander Zien "Semi-Supervised Learning", The MIT Press, 2006.



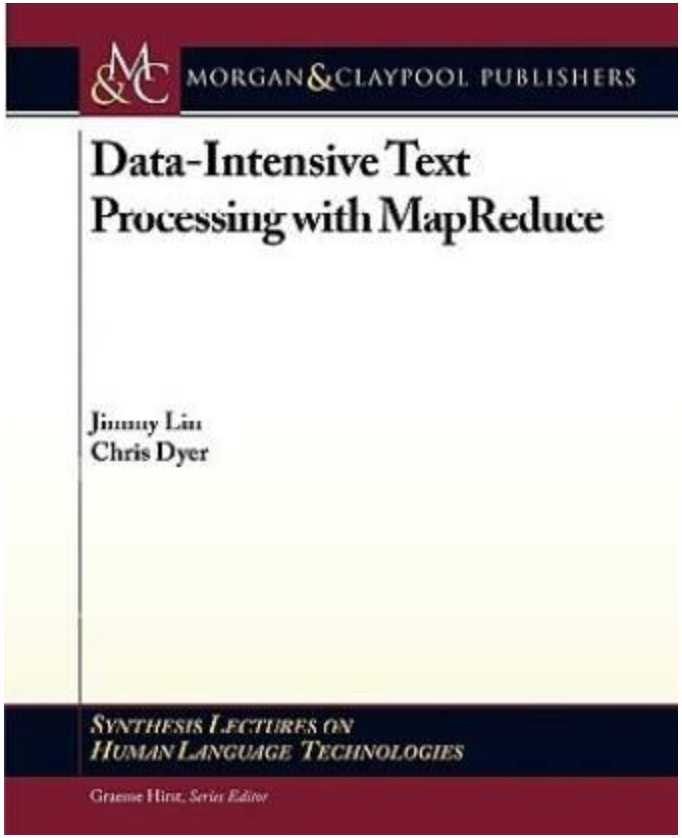
Reference Books

- Ron Bekkerman, Mikhail Bilenko, John Langford "Scaling Up Machine Learning: Parallel and Distributed Approaches", Cambridge University Press, 2012.



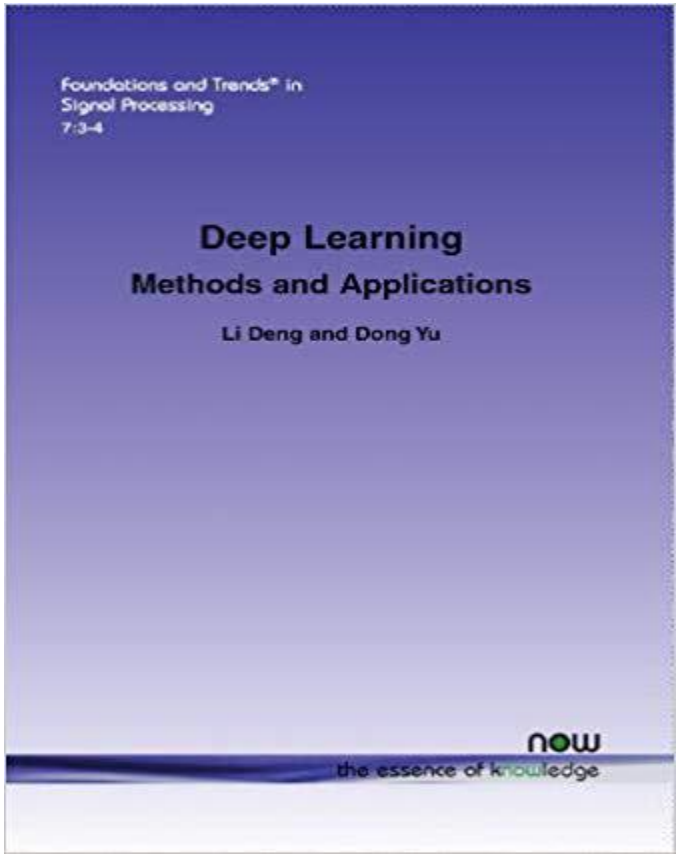
Reference Books

- Jimmy Lin, Chris Dyer, "Data-Intensive Text Processing with MapReduce", Morgan & Claypool Publishers, 2010.



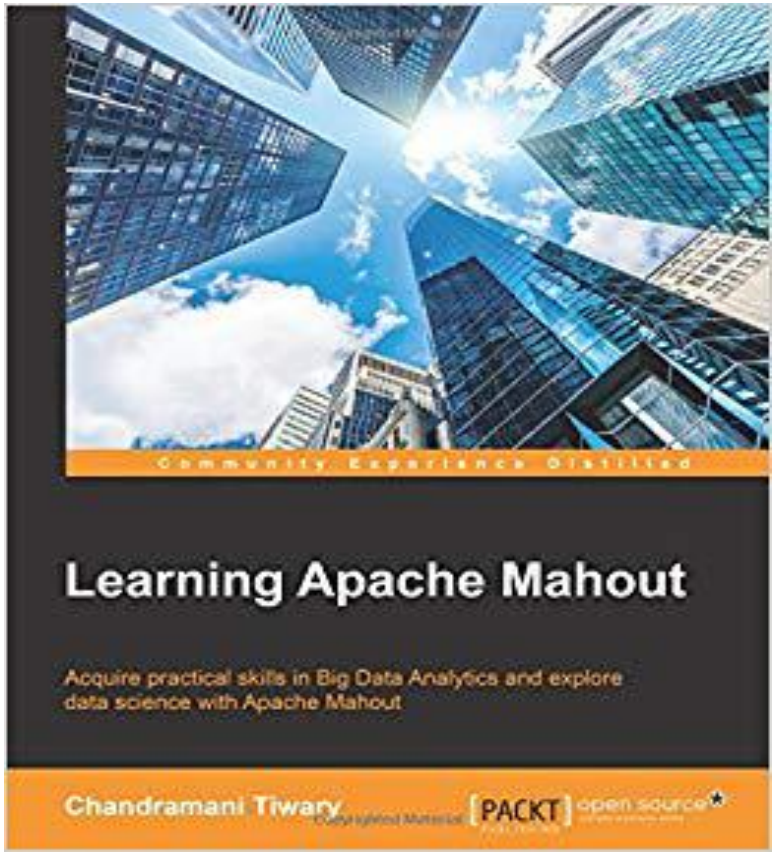
Reference Books

- Li Deng, Dong Yu, “Deep Learning: Methods and Applications”, Now Publisher, 2014.



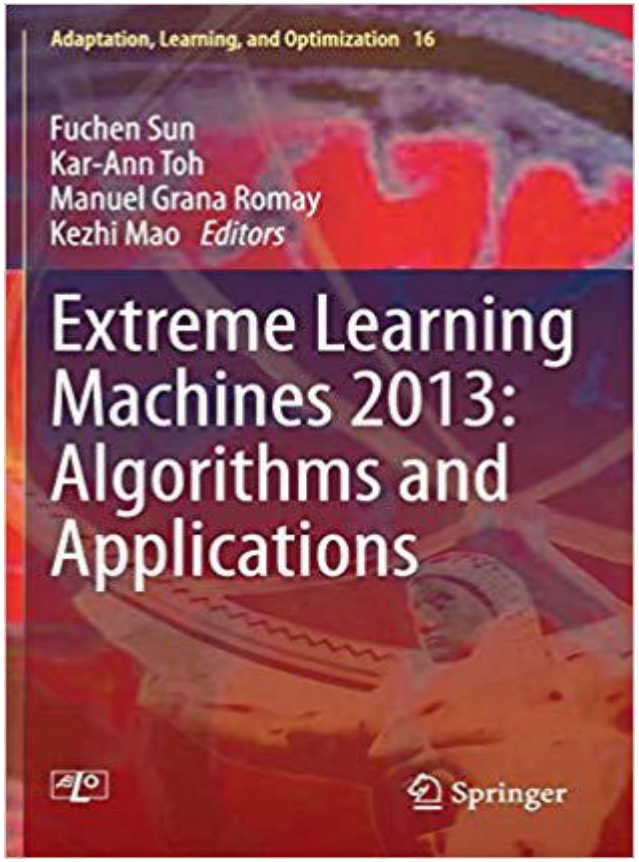
Reference Books

- ChandramaniTiwary "Learning Apache Mahout", Packt Publishing, 2015.



Reference Books

- Fuchen Sun, Kar-Ann Toh, Manuel Grana Romay, KezhiMao,"Extreme Learning Machines 2013: Algorithms and Applications", Springer, 2014.



List of Experiments

- 1. K-means implementation in MapReduce
- 2. Association Rule Mining with MapReduce
- 3. Decision trees in Spark and MapReduce
- 4. Naïve Bayes classification using Spark and MapReduce
- 5. Advanced text processing with Spark
- 6. Clustering models with Spark
- 7. Building a recommendation engine with Spark

List of Experiments

- 8. Representing social-network data using Graphs
- 9. Implementing Semi-supervised Clustering
- 10. Deep Learning using H_2O
- 11. Predictive analysis using H_2O tool
- 12. SVM Classification using Mahout
- 13. Spectral clustering using Mahout
- 14. Building a recommendation engine with Sparkling water
- 15. Deep Learning using DL4J