



Mu Sigma

## Introduction to Mahout and Java

*Do The Math*

Chicago, IL  
Bangalore, India  
[www.mu-sigma.com](http://www.mu-sigma.com)

February 1, 2012

Proprietary Information

"This document and its attachments are confidential. Any unauthorized copying, disclosure or distribution of the material is strictly forbidden"

# Agenda

- ▶ Machine Learning
  - Introduction
  - Types
  - Use Cases
  
- ▶ Mahout
  - Introduction
  - Themes
  - A few algorithms
  - Command line usage
  
- ▶ Exercises
  
- ▶ Appendix

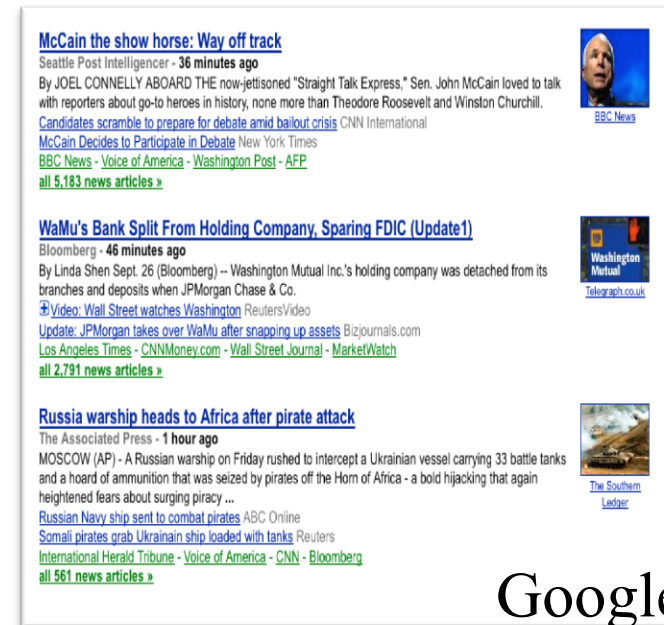


# What is Machine Learning?

- ▶ “Machine Learning is programming computers to optimize a performance criterion using example data or past experience”
  - *Intro. To Machine Learning* by E. Alpaydin
- ▶ This branch of AI helps in recognizing patterns and make intelligent decisions based on known characteristics
- ▶ Some common characteristics of usage:
  - Used when dealing with large volumes of data
  - There must be identifiable features in the dataset
  - Last but not the least, the data is too big or costly for people to handle (people can still help though by creating a training dataset)

# Machine Learning: What are the different types of algorithms?

- ▶ Supervised Learning
  - Using labeled training data, create function that predicts output of non-familiar inputs
- ▶ Unsupervised Learning
  - Use unlabeled data, create function that predicts output
- ▶ Semi-supervised Learning
  - Use labeled and unlabeled data

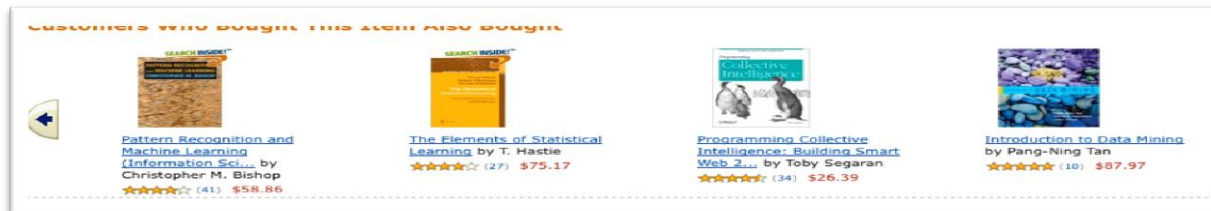


**McCain the show horse: Way off track**  
Seattle Post Intelligencer - 36 minutes ago  
By JOEL CONNELLY ABOARD THE now-jetisoned "Straight Talk Express," Sen. John McCain loved to talk with reporters about go-to heroes in history, none more than Theodore Roosevelt and Winston Churchill.  
[Candidates scramble to prepare for debate amid bailout crisis](#) CNN International  
[McCain Decides to Participate in Debate](#) New York Times  
[BBC News - Voice of America - Washington Post - AFP](#)  
[all 5,183 news articles »](#)

**WaMu's Bank Split From Holding Company, Sparing FDIC (Update)**  
Bloomberg - 46 minutes ago  
By Linda Shen Sept. 26 (Bloomberg) -- Washington Mutual Inc.'s holding company was detached from its branches and deposits when JPMorgan Chase & Co.  
[Video: Wall Street watches Washington](#) Reuters/Video  
Update: JPMorgan takes over WaMu after snapping up assets Bizjournals.com  
[Los Angeles Times - CNNMoney.com - Wall Street Journal - MarketWatch](#)  
[all 2,791 news articles »](#)

**Russia warship heads to Africa after pirate attack**  
The Associated Press - 1 hour ago  
MOSCOW (AP) - A Russian warship on Friday rushed to intercept a Ukrainian vessel carrying 33 battle tanks and a hoard of ammunition that was seized by pirates off the Horn of Africa - a bold hijacking that again heightened fears about surging piracy ...  
[Russian Navy ship sent to combat pirates](#) ABC Online  
[Somali pirates grab Ukrainian ship loaded with tanks](#) Reuters  
[International Herald Tribune - Voice of America - CNN - Bloomberg](#)  
[all 561 news articles »](#)

Google News



Customers who bought this item also bought

Book Title	Author	Rating	Price
Pattern Recognition and Machine Learning	Christopher M. Bishop	★★★★☆ (41)	\$58.86
The Elements of Statistical Learning	T. Hastie	★★★★☆ (27)	\$75.17
Programming Collective Intelligence: Building Smart Web 2.0	Toby Segaran	★★★★☆ (34)	\$26.39
Introduction to Data Mining	Pang-Ning Tan	★★★★☆ (10)	\$87.97

Amazon.com



## Machine Learning: Different use cases

- ▶ Recommend friends/dates/products
- ▶ Classify content into predefined groups
- ▶ Find similar content based on object properties
- ▶ Find associations/patterns in actions/behaviors
- ▶ Identify key topics in large collections of text
- ▶ Detect anomalies in machine output
- ▶ Ranking search results
- ▶ Others?

## Agenda

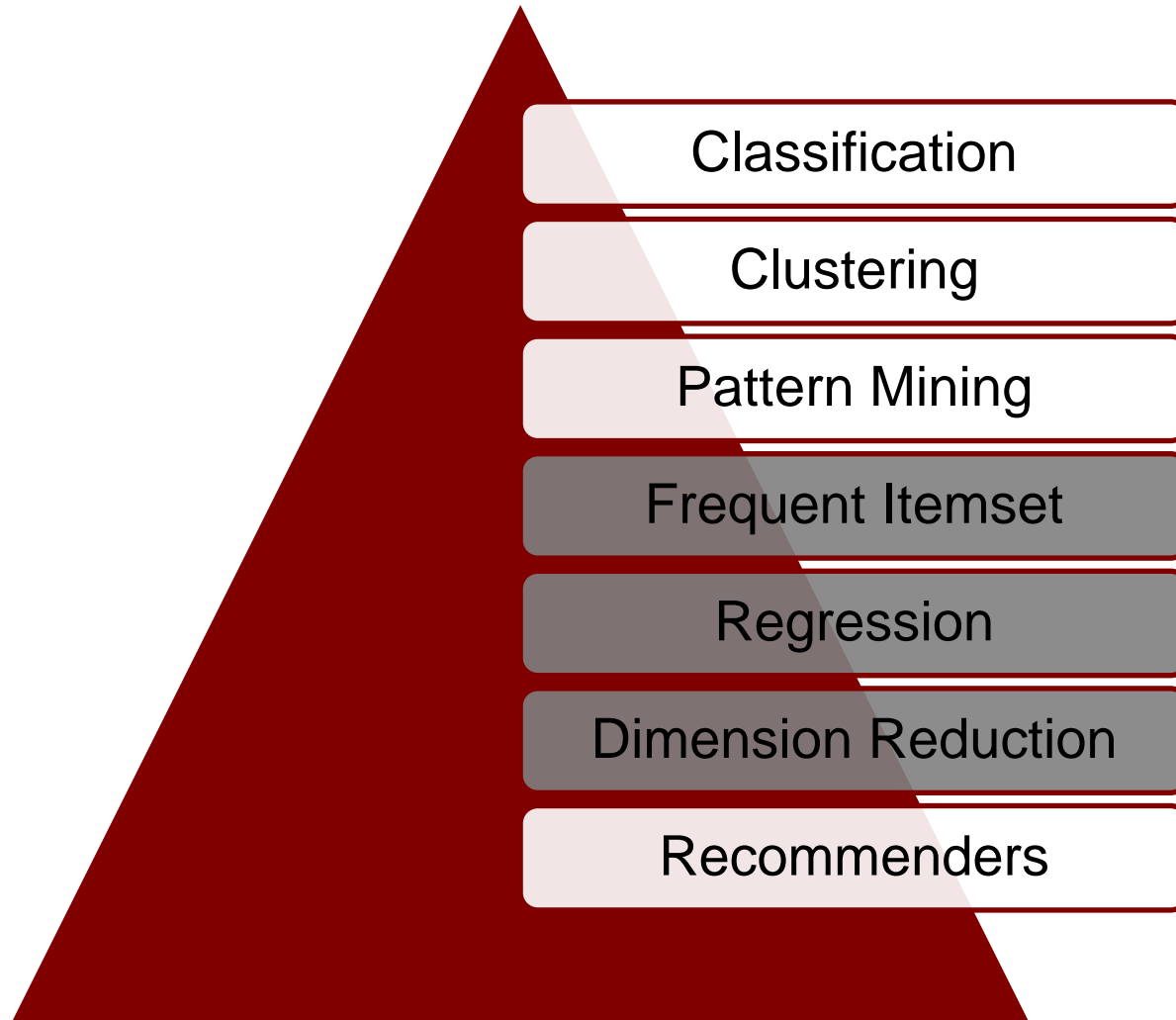
- ▶ Machine Learning
  - Introduction
  - Types
  - Use Cases
- ▶ Mahout
  - Introduction
  - Themes
  - A few algorithms
  - Command line usage
- ▶ Exercises
- ▶ Appendix

## What is Mahout?



- ▶ Apache Software Foundation project to create
  - scalable machine learning libraries
  - Apache Software License(open source and commercially free to use)
- ▶ Many open source Machine Learning libraries lack:
  - A good community
  - Documentation and Examples
  - Scalability
  - Or are completely research oriented
- ▶ Is NOT an execution environment but a library of machine learning algorithms that run on top of Hadoop
- ▶ Tremendous growth of the project, is only a few years old
- ▶ Intelligent Apps are the Present and Future

## Mahout: What are different types of Algorithms?







## Mahout: What are classification algorithms?

- ▶ Label previously unseen objects so as to group them together
- ▶ Example
  - For text data, one could assign customer complaints to LOB's
- ▶ Examples:
  - Spam Filtering
  - Named Entity Recognition
  - Phrase Identification
  - Sentiment Analysis
  - Classification into a Taxonomy

Logistic Regression

Bayesian

Support Vector Machines

Neural Network

Random Forest

Online Passive Aggressive

Boosting

Hidden Markov Models

## Mahout: What are recommenders?

- ▶ Set of algorithms which provide recommendations based on properties of objects
- ▶ Extensive framework for collaborative filtering
- ▶ Recommenders
  - User based
  - Item based

Non-distributed Recommenders

Distributed Item-based Collaborative Filtering

Collaborative Filtering using Parallel Matrix Factorization



Nikon D90 12.3MP DX-Format CMOS Digital SLR Camera with 18-105 mm f/3.5-5.6G ED AF-S VR DX Nikkor Zoom Lens  
by Nikon

★★★★☆ (561 customer reviews) | Like (209)

List Price: ~~\$4,199.95~~  
Price: [See price in cart](#) (Why don't we show the price?)

15 new    25 used from \$850.00    7 refurbished from \$1,049.99

Style: With 18-105mm Lens

[D90 Body Only](#)    [With 18-105mm Lens](#)

### The Amazon recommendation engine

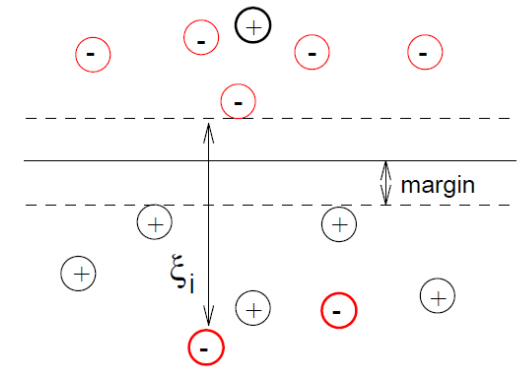
Page 1 of 16

**Customers Who Bought This Item Also Bought**

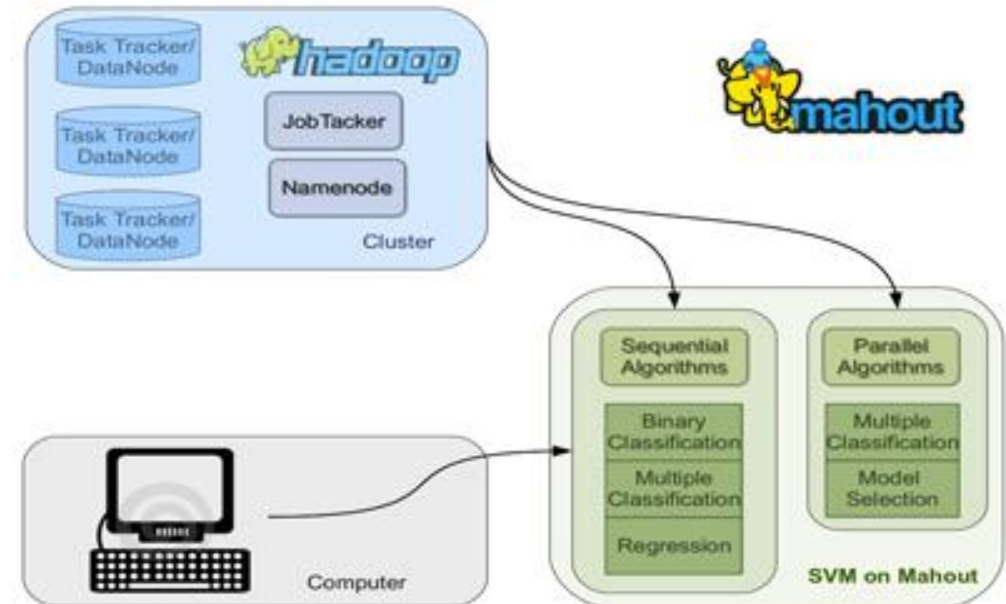
 <p>Transcend 16 GB SDHC Class 10 Flash Memory Card TS16GSDHC10E ★★★★☆ (1,504) \$18.94</p>	 <p>Case Logic SLRC-201 SLR Zoom Holster (Black) ★★★★☆ (162) \$26.91</p>	 <p>Tiffen 67mm UV Protection Filter ★★★★☆ (1,015) \$12.96</p>	 <p>16GB SDHC HC-SD MEMORY CARD FOR NIKON CAMERA SLR D90 by Patriot ★★★★☆ (6) \$15.90</p>	 <p>Nikon D90 For Dummies by Julie Adair King ★★★★☆ (46) \$19.79</p>	 <p>Mastering the Nikon D90 by Darrell Young ★★★★☆ (37) \$23.07</p>
---	---	---	---	---	--

# Mahout: Discussing the current state of Support Vector Machines

- ▶ Generation of learning algorithms that is used to solve binary classifications or regressions
- ▶ Considers objects as points in an n-dimensional feature space, each object is assigned a binary label(positive or negative)
- ▶ Many variations, Sequential SVM solver based on the Pegasos\* algorithm for Primal SVM is implemented as a patch but that doesn't really help
- ▶ [Mahout 14](#), [Mahout 232](#), [Mahout 334](#)

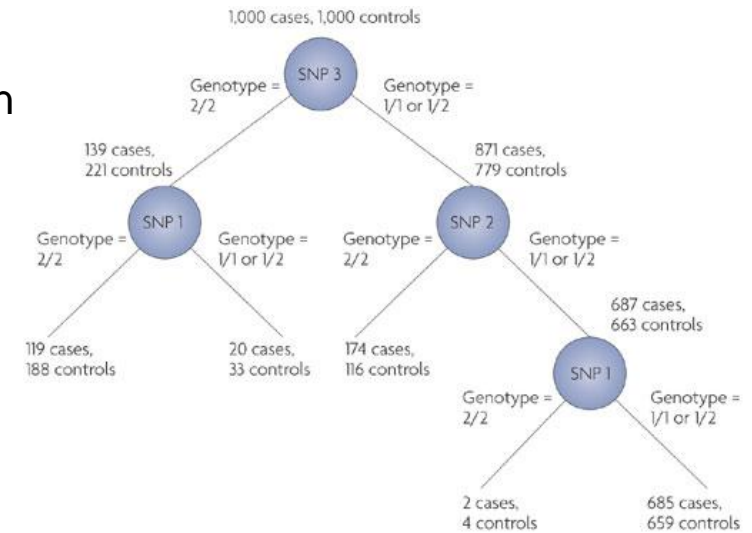


Still awaiting a merge, performance pretty much the same as an algorithm in R



## Mahout: Discussing the current state of Random Forest

- ▶ Developed by Leo Breiman and Adele Cutler, is an *ensemble classifier* that consists of growing many decision trees
- ▶ One of the more accurate learning algorithms
- ▶ Needs number of trees to be used and the number of variables ( $m$ ) to be randomly selected from the available set of variables as input
- ▶ In-memory and partial implementation available at <https://cwiki.apache.org/MAHOUT/partial-implementation.html>



Nature Reviews | Genetics

Source: Nature.com

No progress has been made in almost two years in [Mahout 145](#), last update - September, 2009



## Mahout: Command line usage

- ▶ Shell script in `$MAHOUT_HOME/bin` helps with most tasks
- ▶ Different algorithms will require different setup, for mahout command line one needs to be aware of the job specific options
  - ▶ Typically, the first thing to do would be to lookup options available in mahout
  - ▶ At the console, type `mahout`, press a tab and then press enter
  - ▶ This brings up a list of valid program names, e.g. `canopy: : Canopy clustering`
- ▶ Each program has an independent list of parameter that it requires, which can be looked up by saying:
  - ▶ `mahout <program name> --help`, e.g. `mahout kmeans --help`
- ▶ Let's try that

## Agenda

- ▶ Machine Learning
  - Introduction
  - Types
  - Use Cases
  
- ▶ Mahout
  - Introduction
  - Themes
  - A few algorithms
  - Command line usage
  
- ▶ Exercises
  
- ▶ Appendix

## Example: K-Means clustering and LDA

- ▶ Using mahout examples, build-reuters.sh(uses the Reuters dataset)
  - With k-means
  - With lda
  - Type, `cd /usr/local/hadoop/mahout/mahout-distribution-0.5/examples/bin/`
  
- ▶ Setup Eclipse
  - Install m2Eclipse
  - Create a project
  - Download source for Mahout 0.5
  - Run the k-means example
    - » Build the jar
  - Run the lda-example
    - » Build the jar



## Agenda

- ▶ Machine Learning
  - Introduction
  - Types
  - Use Cases
  
- ▶ Mahout
  - Introduction
  - Themes
  - A few algorithms
  - Command line usage
  
- ▶ Exercises
  
- ▶ Appendix



## Related links

- ▶ Mahout main page: <http://mahout.apache.org/>
- ▶ The Mahout cwiki: <https://cwiki.apache.org/MAHOUT/mahout-wiki.html>
- ▶ Mahout in Action example source code: [http://manning.com/owen/MiA\\_SourceCode.zip](http://manning.com/owen/MiA_SourceCode.zip)



**Thank You**

**Chicago, IL  
Bangalore, India  
February 1, 2012  
[www.mu-sigma.com](http://www.mu-sigma.com)**

Proprietary Information

"This document and its attachments are confidential. Any unauthorized copying, disclosure or distribution of the material is strictly prohibited"