ApacheCon Europe '08

DRAFT

MAHOUT

Apache Mahout

Bringing Machine Learning to Industrial Strength

Presented by: Isabel Drost
(neofonie GmbH)

Problem Setting

DRAFT

MAHOUT

- Huge amounts of data at our fingertips.

Web pages

Mail archives

Pictures tagged with topics

Search engine logs

Wiki collaboration data

Source control logs

Information on proteins

News articles

Corporate job postings

Social network graphs

Traffic data

Tagged and rated videos

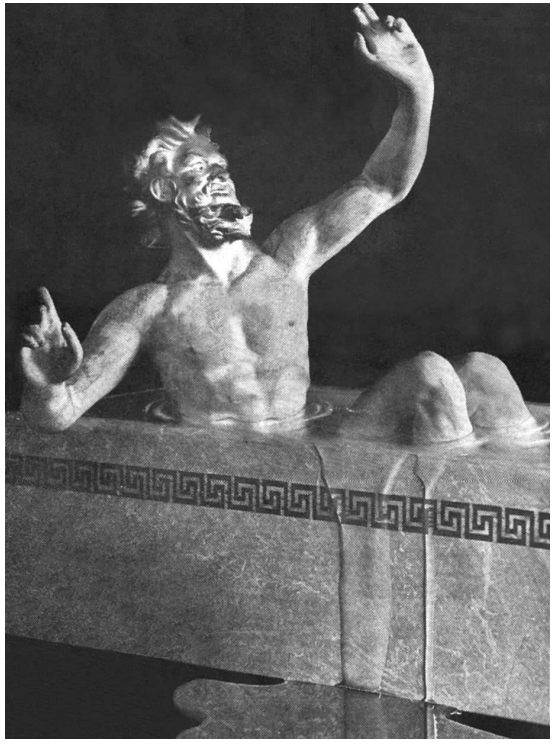
- Need means to deal with all the data.

Problem Setting

DRAFT

MAHOUT

- Nature generates data.



- Archimedes generates model.

$$\frac{Density\ of\ Object}{Density\ of\ Fluid} = .$$

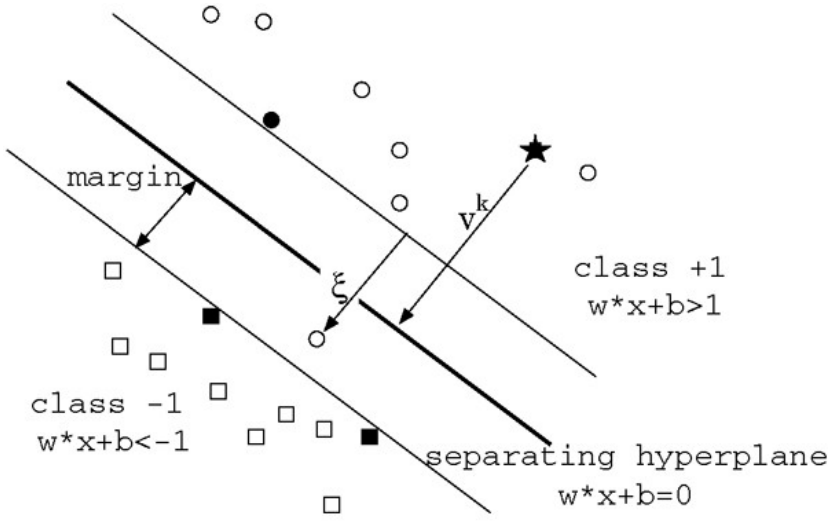
$$\frac{Weight}{Weight - Apparent\ immersed\ weight}$$

Problem Setting

DRAFT

MAHOUT

- Nature generates data.
- ML generates models.

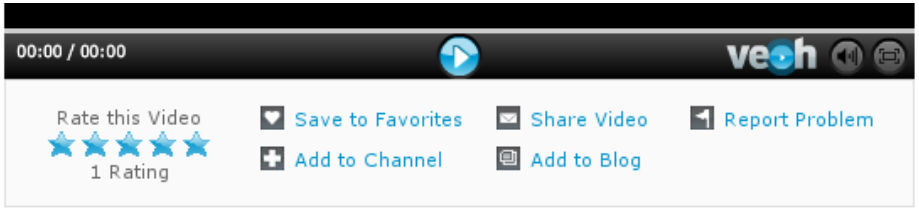


Where is ML used already?

DRAFT

MAHOUT

- Search result clustering.
- “Did you mean” feature.
- Auto completion.
- Language detection.
- Analysis of tags.



Once upon a time

- How it all began:
  - Summer 2007: Crazy developers needed scalable ML.
  - Mailing list and wiki followed quickly.
- Contacted people
  - from research.
  - from related Apache projects.
- Rather large community even before project start.
- 25.01.2008: Project Mahout launched.

Who we are

DRAFT

MAHOUT

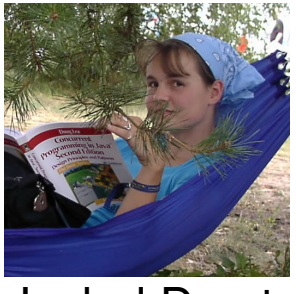
Dawid Weiss
Carrot2

Karl Wettin
Lucene

Grant Ingersoll
Lucene PMC

Ted Dunning
The Veoh guy

Jeff Eastman
Welcome!



Isabel Drost
(that would be myself)

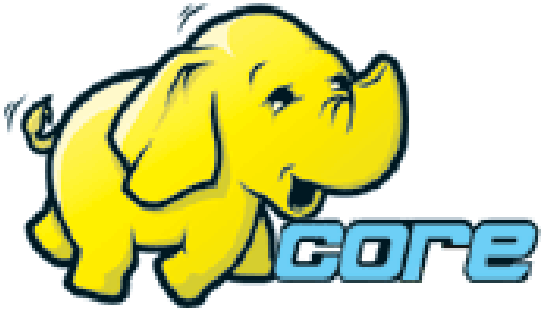
Otis Gospodetnic – Lucene
Erik Hatcher – Lucene (among others)

Our Mission

DRAFT

MAHOUT

- Build learning algorithms that are scalable.
- Context:



Hadoop – parallelization

Hama – matrix support

Lucene – provides the use cases

Initial contributions

- k-Means implementation.
  - Started with non-parallel version.
  - Ported to Hadoop already.
- Matrix computation package.
  - Building block of many machine learning algorithms.
  - Together with Hama towards parallel matrices.

Initial Contributions

- Work on Naïve Bayes, Perceptron, PLSI/EM.
- Integrate Taste
  - Collaborative filtering project at sourceforge.
  - Item based recommendation.
  - User based recommendation.

GSoC @ Mahout

- *"Implementing Logistic Regression in Mahout"*
- *"Codename Mahout.GA for mahout-machine-learning"*
- *"The Implementation of Support Vector Machine Algorithm at Hadoop Platform"*
- *"Application to participate in Mahout"*
- *"Mahout application (Neural Networks)"*
- *"DeCoDe - A smart code search engine based on lucene to show how Mahout work."*
- *"Applying for mahout machine learning (Neural Networks)"*
- *"Implementation of the Principal Compenents Analysis algorithm for Mahout"*
- *"MAHOUT Naive Bayes implementation"*

Conclusions

- This is just the beginning: Even the logo is a draft :)
- High demand for scalable machine learning.
- We need you – in case you have:
  - A good deal of enthusiasm.
  - Solid mathematical knowledge to understand ML papers.
  - Either proficient in or willing to learn about Hadoop.
  - Or: A lot of data and want to know what to learn from it.
- mahout-dev@apache.org mahout-user@apache.org