Scalable Machine Learning with Hadoop (most of the time)

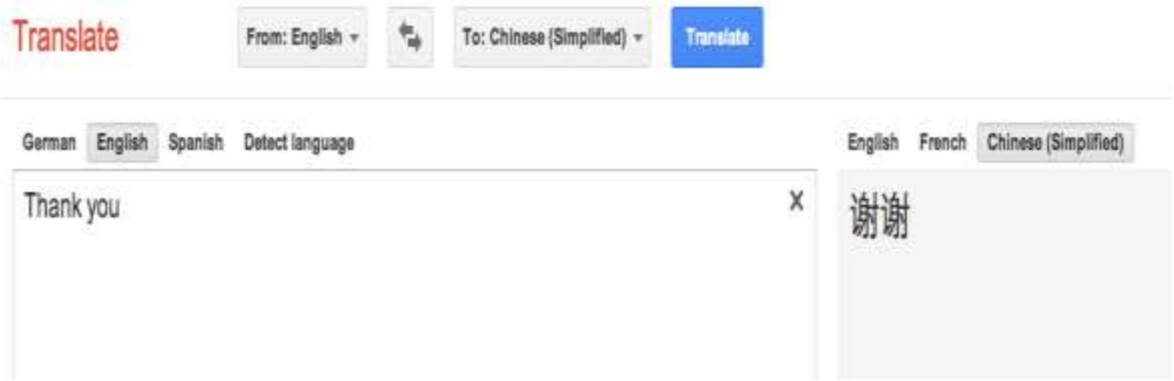
Grant Ingersoll
Chief Scientist

October 2, 2012

Search | Discover | Analyze

Anyone Here Use Machine Learning?

- Any users of:
  - Google?
    - Search
    - Translation
    - Priority Inbox
  - Facebook?
  - Twitter?
  - LinkedIn?



Thank you

谢谢

Google Translate

Topics

- What is scalable machine learning?
- Use Cases
- Approaches
  - Hadoop-based
  - Alternatives
- What is Apache Mahout?

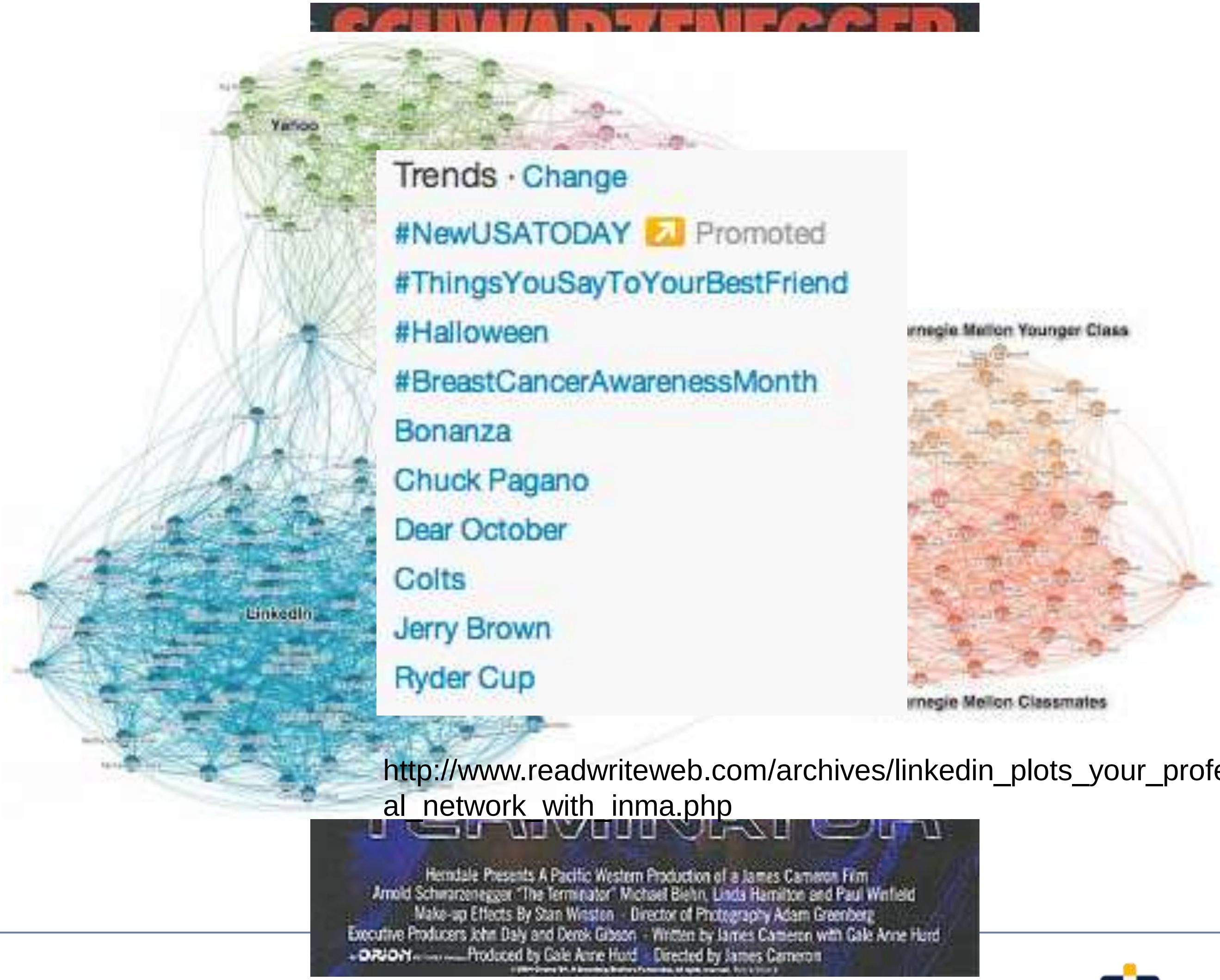
Machine Learning

- “Machine Learning is programming computers to optimize a performance criterion using example data or past experience”
  - *Intro. To Machine Learning* by E. Alpaydin
- Lots of related fields:
  - Information Retrieval
  - Stats
  - Biology
  - Linear algebra
  - Many more

What does scalable mean for us?

- As data grows linearly, either scale linearly in time or in machines
  - 2X data requires 2X time or 2X machines (or less!)
- Goal: Be as fast and efficient as possible given the intrinsic design of the algorithm
  - Some algorithms won't scale to massive machine clusters
  - Others fit logically on a Map Reduce framework like Apache Hadoop
  - Still others will need different distributed programming models
  - Be pragmatic

Common Use Cases



Trends · Change

#NewUSATODAY Promoted

#ThingsYouSayToYourBestFriend

#Halloween

#BreastCancerAwarenessMonth

Bonanza

Chuck Pagano

Dear October

Colts

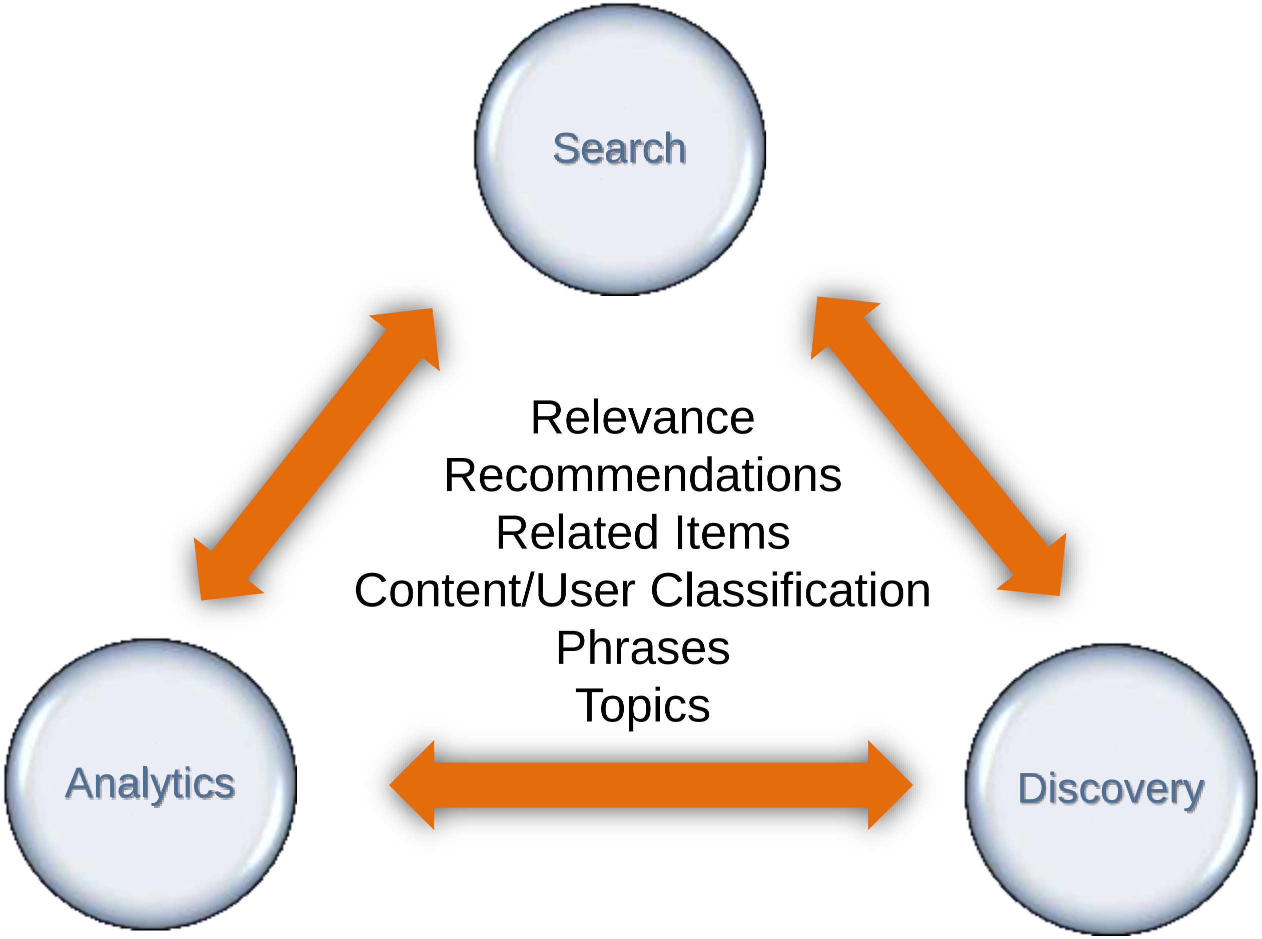
Jerry Brown

Ryder Cup

Yahoo
LinkedIn
rnegie Mellon Younger Class
rnegie Mellon Classmates
Hemdale Presents A Pacific Western Production of a James Cameron Film
Arnold Schwarzenegger "The Terminator" Michael Biehn, Linda Hamilton and Paul Winfield
Make-up Effects By Stan Winston · Director of Photography Adam Greenberg
Executive Producers John Daly and Derek Gibson · Written by James Cameron with Gale Anne Hurd
ORION Produced by Gale Anne Hurd · Directed by James Cameron

http://www.readwriteweb.com/archives/linkedin_plots_your_professional_network_with_inma.php

My Use Cases



Relevance
Recommendations
Related Items
Content/User Classification
Phrases
Topics

Proprietary

Scalable Approaches

- Mind the Gap
  - Algorithms are the fun stuff, but you'll spend more time on ETL, feature selection and post-processing
  - Simpler is usually better at scale

1. Scale Data Pipeline -> Sample -> Sequential
2. Hadoop
3. Ensemble (distribute many sequential models)
4. Spark, MPI & BSP, Others

Open Source Machine Learning Libraries

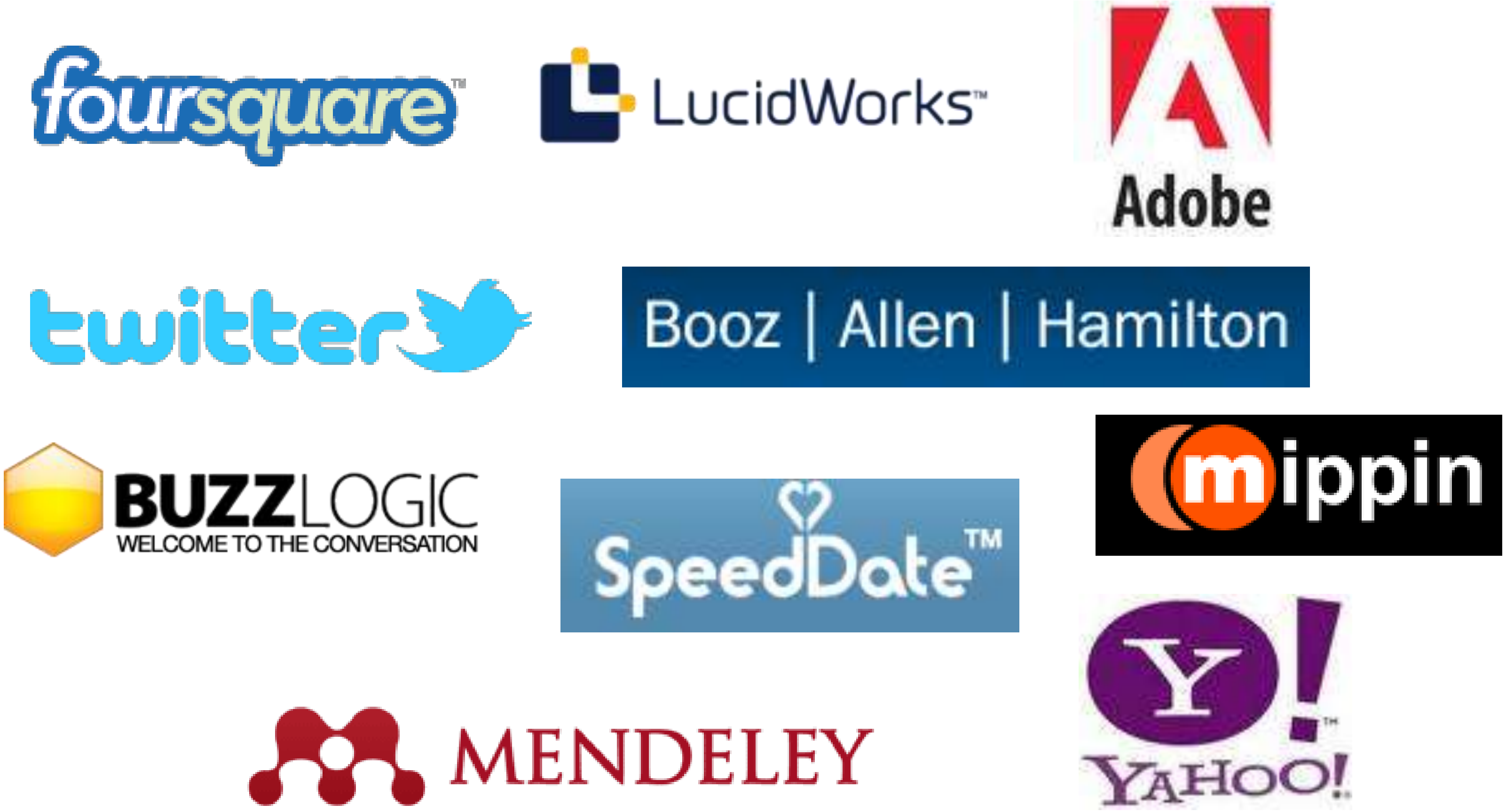
- Apache Mahout
- Vowpal Wabbit
- R Stats Project
- Weka
- LibSVM, SVMLight
- Many, many more

Apache Mahout

- An Apache Software Foundation project to create scalable machine learning libraries under the Apache Software License
  - http://mahout.apache.org
- Why Mahout?
  - Many Open Source ML libraries are either:
    - Lack Community
    - Lack Documentation and Examples
    - Lack Scalability
    - Lack the Apache License
    - Or are research-oriented

http://dictionary.reference.com/browse/mahout

Who uses Mahout?



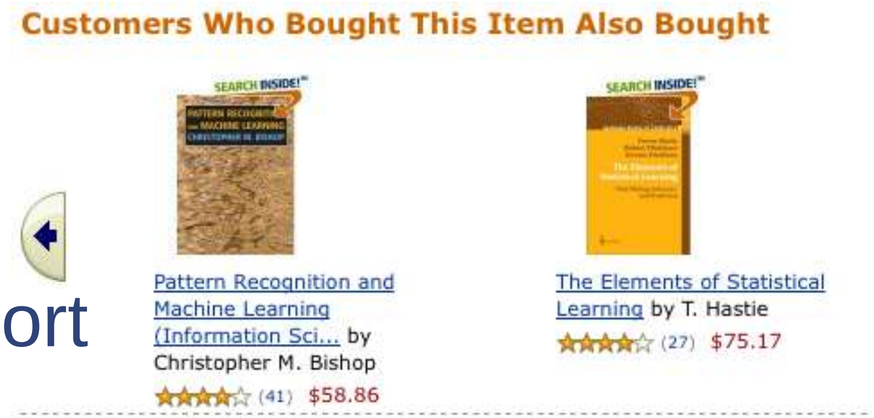
https://cwiki.apache.org/confluence/display/MAHOUT/Powered+By+Mahout

What Can I do with Mahout Right Now?

3 “C”s + Extras

Collaborative Filtering

- Recommender Approaches
  - User based
  - Item based



Customers Who Bought This Item Also Bought

Pattern Recognition and Machine Learning (Information Sci... by Christopher M. Bishop
★★★★☆ (41) $58.86

The Elements of Statistical Learning by T. Hastie
★★★★☆ (27) $75.17

- Online and Offline support
  - Offline can utilize Hadoop

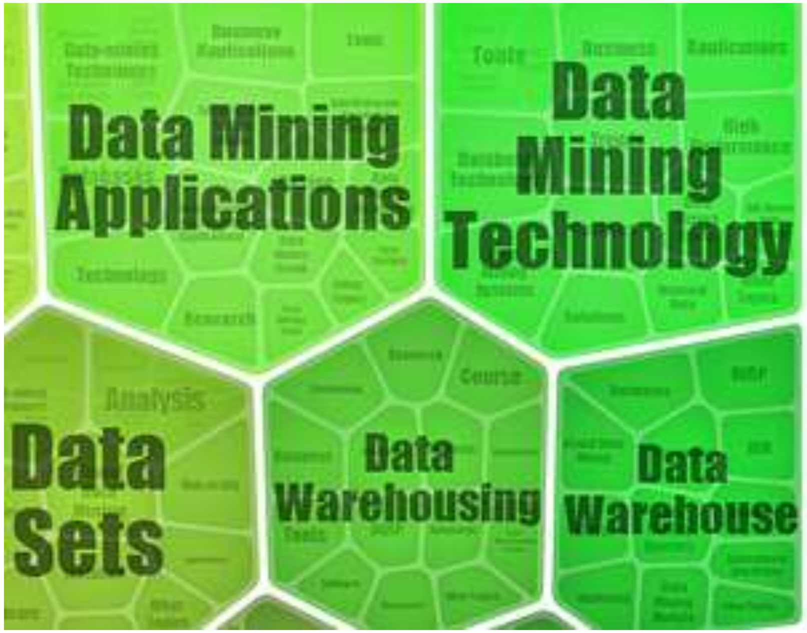
- Many different Similarity measures
  - Cosine, LLR, Tanimoto, Pearson, others

Hadoop Recommenders

- Alternating Least Squares
  - Iterative, but scales well
  - Deals well with sparseness
  - "Large-scale Parallel Collaborative Filtering for the Netflix Prize" by Zhou et. al
  - https://cwiki.apache.org/MAHOUT/collaborative-filtering-with-als-wr.html
- Slope One
  - Simple yet effective
- Pseudo
  - Distribute sequential approach across Hadoop nodes

Clustering

- Document level
  - Group documents based on a notion of similarity
  - K-Means, Fuzzy K-Means, Dirichlet, Canopy, Mean-Shift, Spectral, Top-Down
  - Pluggable Distance Measures



http://carrotsearch.com/foamtree-overview.html

- Topic Modeling
  - Cluster words across documents to identify topics
  - Latent Dirichlet Allocation
    - Using Collapsed Variational Bayes

Clustering In Hadoop

- Many people start with K-Means, but others can be more effective
- Challenges
  - Iterative nature of many clustering algorithms can be slow
  - Distance measures and other factors can have dramatic impact on performance and quality
  - When in doubt, experiment

Classification

- Place new items into predefined categories
- Online and Offline supported

Shop It To Me

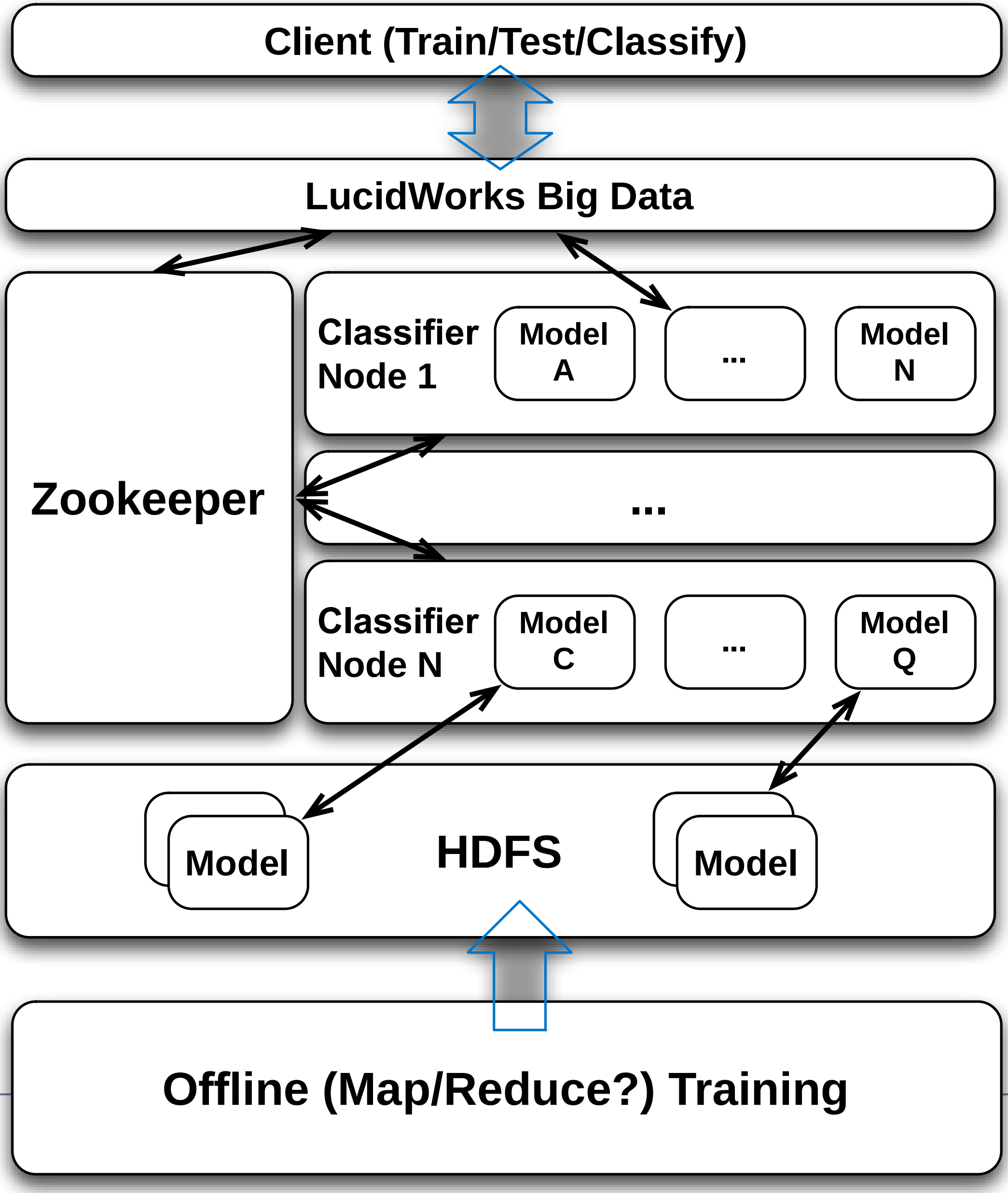
"This gives a raw classification rate requirement of tens of millions of classifications per second, which is, as they say in the old country, a lot."

"Mahout in Action" http://awe.sm/5FyNe

- Hadoop
  - Naïve Bayes
  - Complementary Naïve Bayes
  - Decision Forests
  - Clustering-based

- Sequential
  - Logistic Regression
    - Stochastic Grad. Descent
  - Hidden Markov Model
  - Winnow/Perceptron

Scaling Mahout Classification



Other Mahout Features

- Apache Licensed:
  - Primitive Collections!
  - Extensive Math library
    - Vectors, Matrices, Statistics, etc.
    - Vector Encoding options
- Singular Value Decomposition
- Frequent Pattern Mining
- Collocations (statistically interesting phrases)
- I/O: Lucene, Cassandra, MongoDB and others

What’s Next for Mahout?

- Streaming K-Means
- Map/Reduce Training for HMM?
- Clean Up towards 1.0 release
- 1.0?

Resources

- http://www.lucidworks.com
- grant@lucidworks.com
- @gsingers