lucid IMAGINATION

Open.
Scalable.
Intelligent?

Thinking Lucene Think Lucid.

Free

Unstructured

Mind

Too

Open Source

Ended

For Business

Unstructured Data

- Some estimate (pre-Twitter!) as much as 85% of all data is unstructured
  - Much of it is text
- How well you deal with unstructured data is often the difference maker for an organization
- Is there really such as thing as “pure” unstructured data?

Cascading

All marks are property of their respective owners

Commodity

Big Data

Scale Free

Scalable

Storage

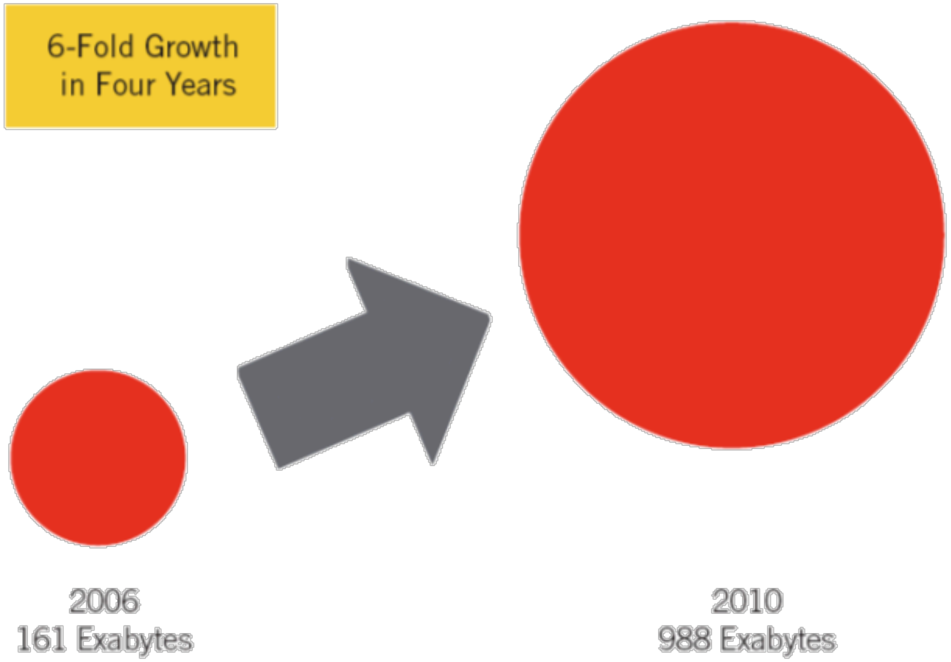
Algorithms

Work force

Fault Tolerant

Distributed

Information Created, Captured and Replicated



http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

We’ve gotten good at…

Data

+

Apache Solr

Lucene

hadoop

and friends

= Open, Scalable Search

The Future is Bright for Scalability

- New Lucene capabilities will give even more control over indexing and searching to allow for exacting control over footprint
- Solr Cloud efforts are integrating ZooKeeper with Solr to make it even easier to manage a large scale Lucene/ Solr installation
  - http://wiki.apache.org/solr/SolrCloud
- Solr + Hadoop makes it easier to index large scale content
  - https://issues.apache.org/jira/browse/SOLR-1301

We've also gotten good at…

Data

+

hadoop

and friends

Proprietary Code

=

Scalable,
Analytics,
Data Crunching,
Social Graph

Organize Discover

Find Associate

Collective

Personalization

Intelligent?

Sentiment Semantics

Learn Plan

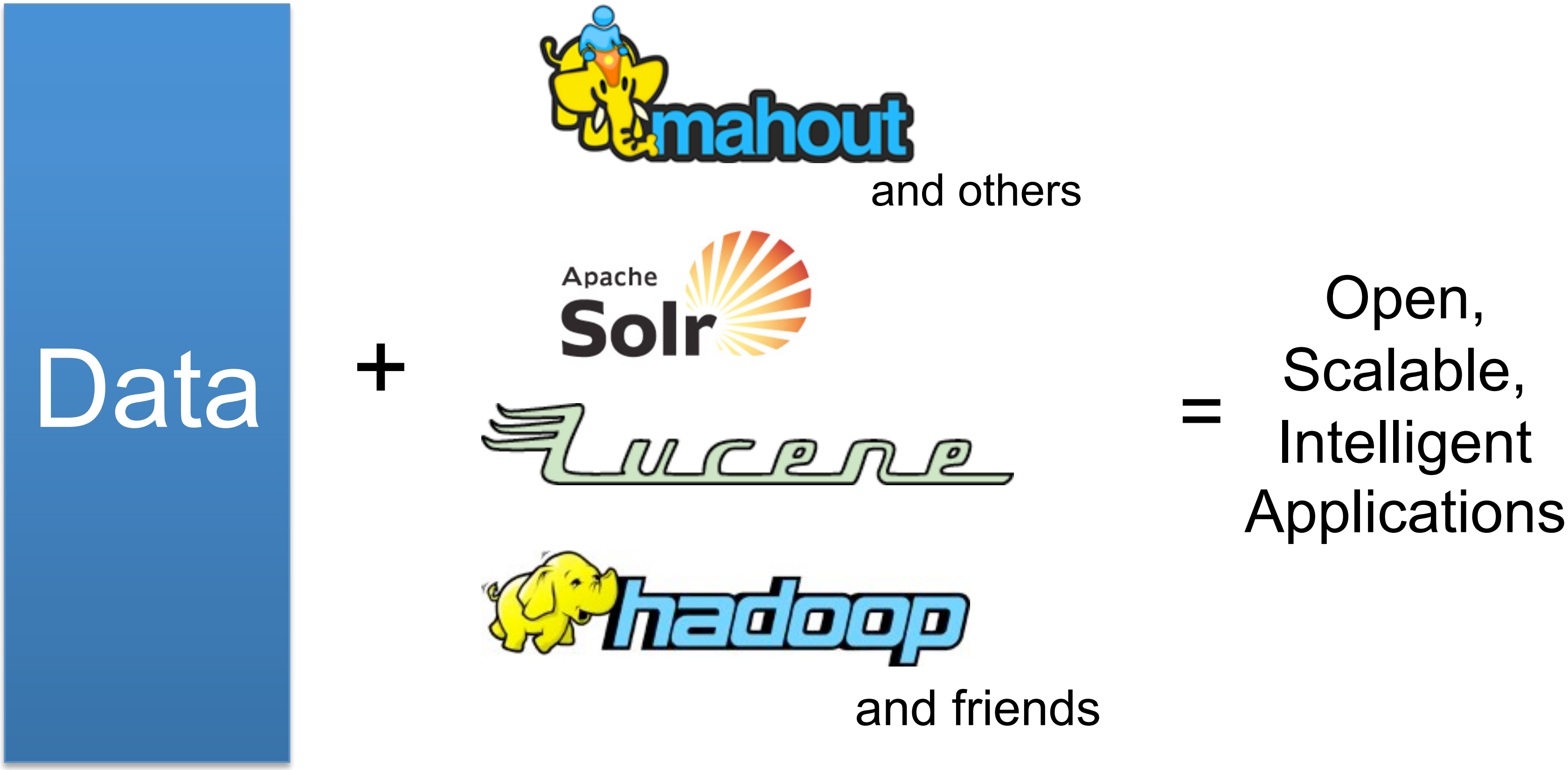
Knowledge

Understand

Reason Solve Problems

Why Should I care?

- Storage, CPU, Memory, Network, Racks, Data Centers, Bandwidth are all commodities
- As are:
    - Search Algorithms
    - Distributed Computing Paradigms
- Open source and scalability demands accelerate commoditization
- Intelligence (artificial and human) is in short supply
- Machine learning can **help**



Data + mahout and others, Apache Solr, Lucene, hadoop and friends = Open, Scalable, Intelligent Applications

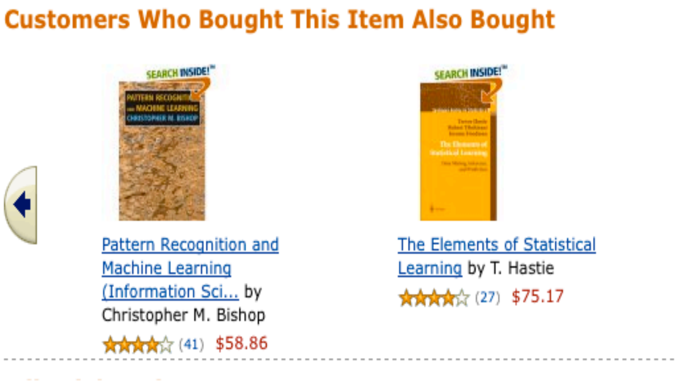
What can you do right now to add intelligence?

Adding Intelligence

- Tip of the Iceberg
- Recommendations
- Organization
- Discovery
- Voice of the Users
- Location Aware
- Make the problem more manageable

Recommendations

- Online and Offline Recommendation capabilities available
  - User-User
  - Item-Item
  - Many different ways to model



Customers Who Bought This Item Also Bought

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

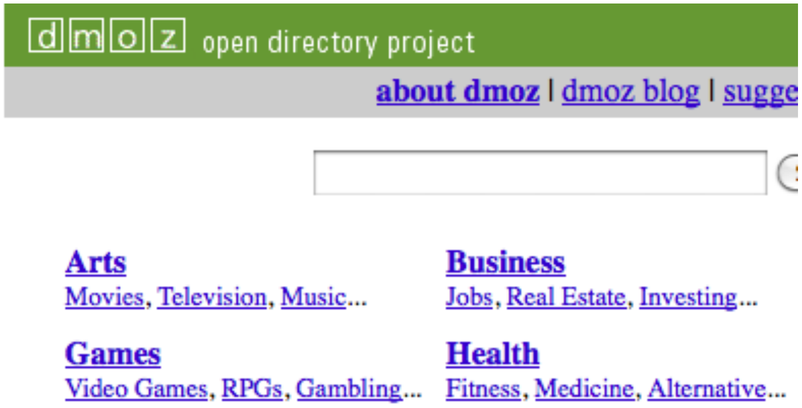
The Elements of Statistical Learning by T. Hastie ★★★★☆ (27) $75.17

- Map/Reduce Ready recommenders available
  - Co-occurrence, pseudo
  - Crude EC2 Estimated Cost: $0.01/1000 recommendations*

* Courtesy Sean Owen

Organization

- Tag/label classify your content into predetermined categories
  - Bayesian and Complementary
  - Random Forests
- Identify Topics
  - Latent Dirichlet Allocation
- All Map/Reduce enabled



dmoz open directory project

about dmoz | dmoz blog | sugge

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Discovery (Mahout)

- Group unseen content via clustering
  - K-Means, Dirichlet, Canopy, etc.
- Frequent Pattern Mining
  - Mine your logs for commonly co-occurring patterns
  - http://www.slideshare.net/hadoopusergroup/mail-antispam
- Collocations
  - Find statistically interesting word co-occurrences (i.e. phrases)
- All Map/Reduce enabled
- http://cwiki.apache.org/MAHOUT/algorithms.html

Discovery (Lucene/Solr)

- Faceting/Drill Downs and other UI summarization
- Auto complete/suggest
  - https://issues.apache.org/jira/browse/SOLR-1316
- Spell Checking
- More Like This and relevance feedback
- Document and Search Result (Carrot^2) clustering

Apache Solr

Share their joys, feel their pain

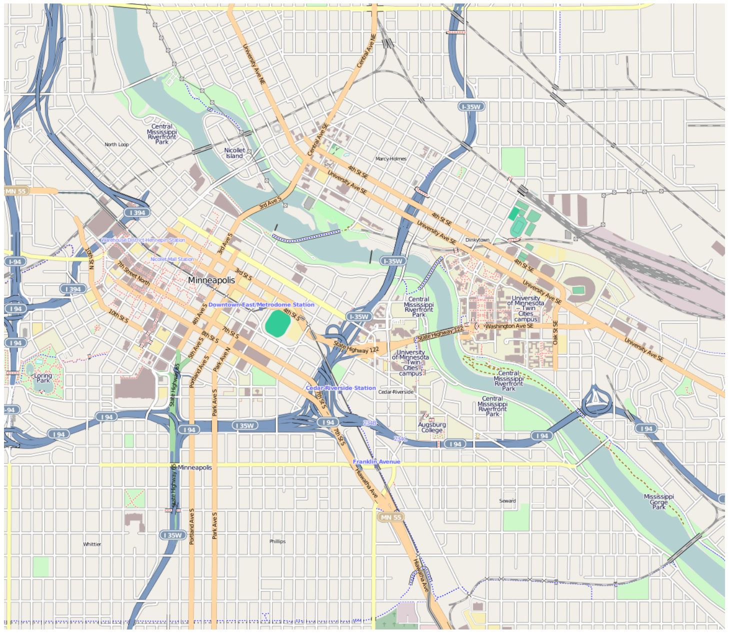
- Understand the voice of the user
- Sentiment Analysis
- Social Network Analysis
- Log Analysis
- Feedback loops

GATE general architecture for text engineering

mahout

Location, Location, Location!

- Providing location aware search results can significantly enhance/reduce the search space for users
- Needs
  - Query Parsing
  - Filtering
  - Boosting
  - Sorting
  - Other



http://www.openstreetmap.org/?lat=44.9744&lon=-93.2484&zoom=14&layers=B000FTFT

Feature Reduction

- Curse of dimensionality!
- Singular Value Decomposition (SVD) is a powerful technique for reducing the dimensionality of large matrices while retaining the core features of the larger space
- Latent Semantic Analysis uses SVD to provide search over the reduced space
  - http://github.com/algoriffic/lsa4solr

Use Case: Enhanced Search

- Latent Semantic Analysis
- Add Collocations or Phrases to your content
- Classify/Cluster your Content
    - Named Entity Recognition, Sentiment analysis, Semantics
    - Facet/Filter
- Related Searches
- Spell Checking
- More Like This
- Clickstream Analysis

openNLP

GATE general architecture for text engineering

Apache Solr

mahout

Where next, Mahout?

- Recommenders
  - Restricted Boltzmann Machines
  - SVD-based
- Classifiers
  - Neural Network
  - Support Vector Machines
  - Stochastic Gradient Descent (logistic regression)
- Clustering
  - Eigen Cuts (spectral clustering)
- Common I/O Formats across algorithms
  - Avro?
- Visualization tools?
- Meta learners?

Open.

Scalable.

Intelligent.

- grant@lucidimagination.com
- @gsingers
- http://www.manning.com/ingersoll