# From Data to Information

## Apache Mahout

Speaker: Isabel Drost

Doug Lea
Concurrent Programming in Java
Second Edition
Design Principles and Patterns

# Isabel Drost

Nighttime:

Came to nutch in 2004.
Co-Founder Apache Mahout.
Organizer of Berlin Hadoop Get Together.

Daytime:

Software developer @ Berlin

Hello FrOSCon visitors!

# Agenda

- Motivation.
- HowTo: A path from data to information.
- Introduction to Mahout.

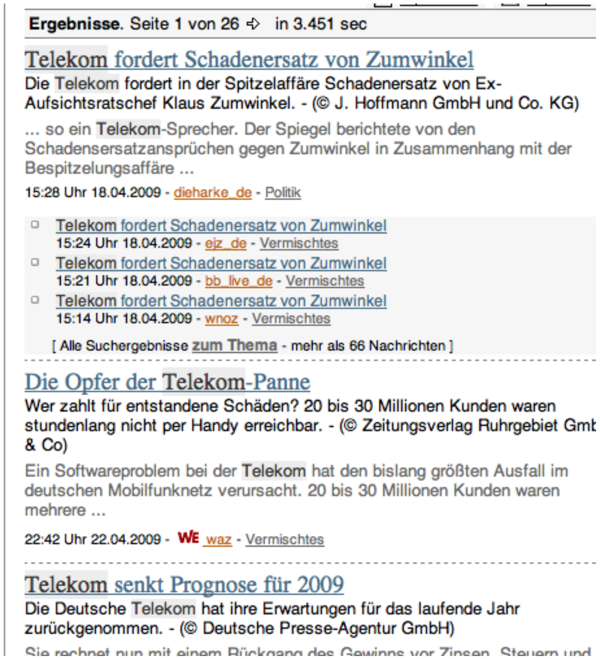
COMMUNITY NEWS
Finishing touches still to come
A glimpse of today, yesterday

January 3, 2006 by Matt Callow
http://www.flickr.com/photos/blackcustard/81680010

# News aggregation

September 10, 2008 by Alex Barth
http://www.flickr.com/photos/a-barth/2846621384

Today: Read news papers, Blogs, Twitter, RSS feed.



**Ergebnisse.** Seite 1 von 26 ➪ in 3.451 sec

Telekom fordert Schadenersatz von Zumwinkel

Die Telekom fordert in der Spitzelaffäre Schadenersatz von Ex-Aufsichtsratschef Klaus Zumwinkel. - (© J. Hoffmann GmbH und Co. KG)

... so ein Telekom-Sprecher. Der Spiegel berichtete von den Schadensersatzansprüchen gegen Zumwinkel in Zusammenhang mit der Bespitzelungsaffäre ...

15:28 Uhr 18.04.2009 - dieharke_de - Politik

- Telekom fordert Schadenersatz von Zumwinkel
  15:24 Uhr 18.04.2009 - ejz_de - Vermischtes
- Telekom fordert Schadenersatz von Zumwinkel
  15:21 Uhr 18.04.2009 - bb_live_de - Vermischtes
- Telekom fordert Schadenersatz von Zumwinkel
  15:14 Uhr 18.04.2009 - wnoz - Vermischtes

[ Alle Suchergebnisse **zum Thema** - mehr als 66 Nachrichten ]

Die Opfer der Telekom-Panne

Wer zahlt für entstandene Schäden? 20 bis 30 Millionen Kunden waren stundenlang nicht per Handy erreichbar. - (© Zeitungsverlag Ruhrgebiet Gmb & Co)

Ein Softwareproblem bei der Telekom hat den bislang größten Ausfall im deutschen Mobilfunknetz verursacht. 20 bis 30 Millionen Kunden waren mehrere ...

22:42 Uhr 22.04.2009 - WE waz - Vermischtes

Telekom senkt Prognose für 2009

Die Deutsche Telekom hat ihre Erwartungen für das laufende Jahr zurückgenommen. - (© Deutsche Presse-Agentur GmbH)

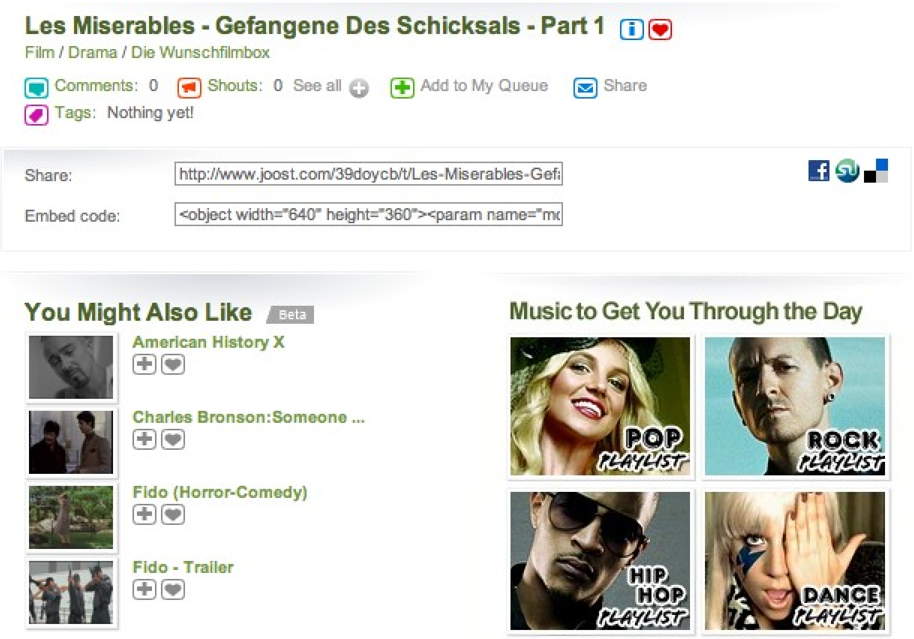
Sie rechnet nun mit einem Rückgang des Gewinns vor Zinsen, Steuern und

Wish: Aggregate sources and track emerging topics.

September 21, 2008, Rodrigo Galindez
http://www.flickr.com/photos/rodrigogalindez/2877367250/

# Go to cinema

March 22, 2008 by Crystian Cruz
http://www.flickr.com/photos/crystiancruz/2353895708



**Les Miserables - Gefangene Des Schicksals - Part 1**

Film / Drama / Die Wunschfilmbox

Comments: 0 Shouts: 0 See all Add to My Queue Share

Tags: Nothing yet!

Share: http://www.joost.com/39doycb/t/Les-Miserables-Gef

Embed code: <object width="640" height="360"><param name="mc

**You Might Also Like** Beta

- American History X
- Charles Bronson:Someone ...
- Fido (Horror-Comedy)
- Fido - Trailer

**Music to Get You Through the Day**

- POP PLAYLIST
- ROCK PLAYLIST
- HIP HOP PLAYLIST
- DANCE PLAYLIST

Today: IMDB, zitty, movie review pages, twitter, blogs, ask friends.

Wish: Reviews, sentiment detection, recommendations.

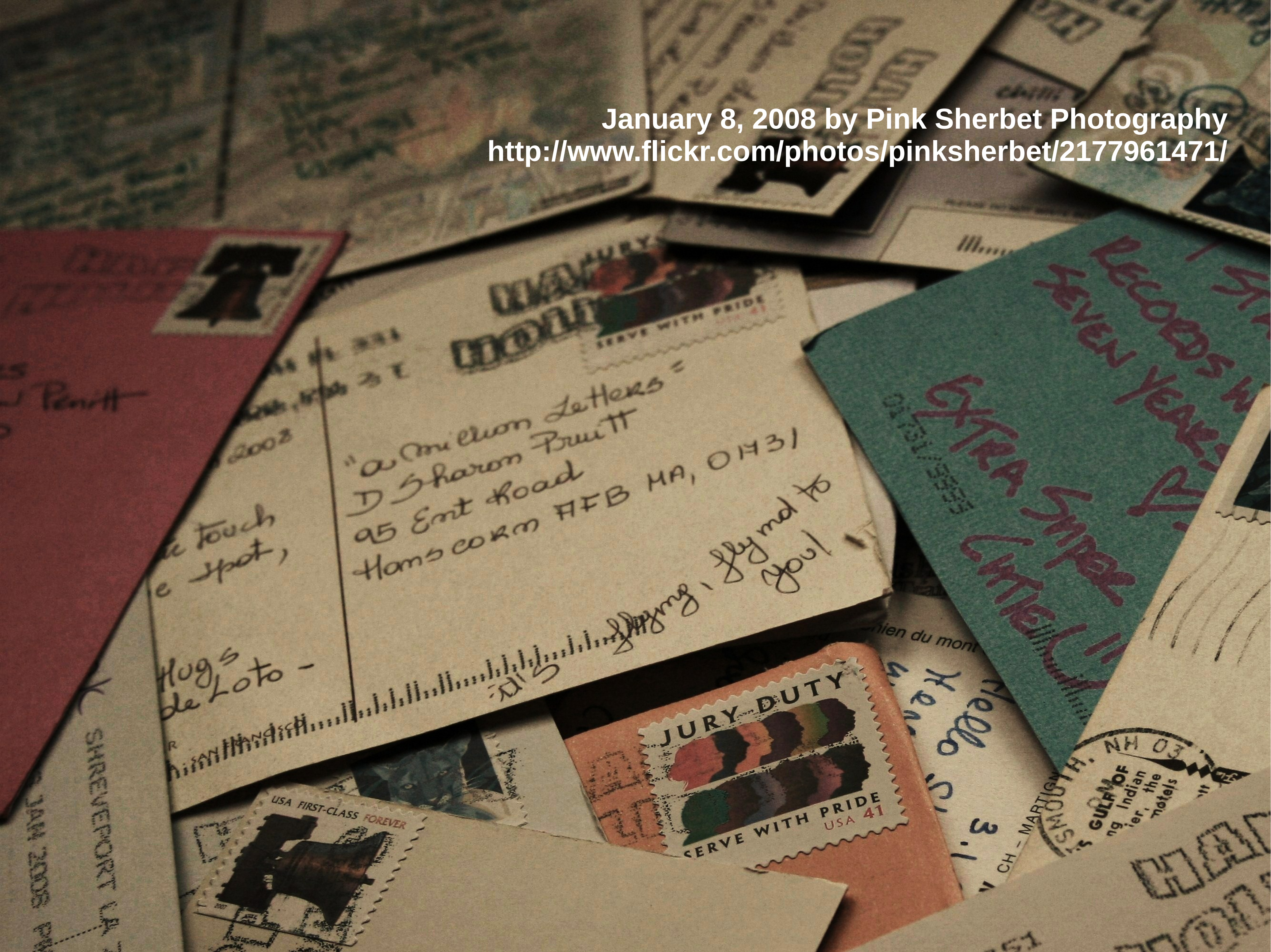
# HowTo: From data to information.

# From data to information.

- Start collecting and storing data.

- Analyse and understand data.

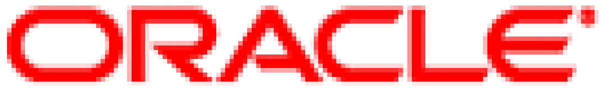
- Answer more complex questions.



January 8, 2008 by Pink Sherbet Photography
http://www.flickr.com/photos/pinksherbet/2177961471/

# Data storage options

- Structured, relational.
  - Customer data.
  - Bug database.


ORACLE

MySQL
Sun
microsystems

PostgreSQL

# Data storage options

- Structured, relational .
  - Customer data.
  - Bug database.

- Continuous files.
  - Log data.
  - Document Stream.

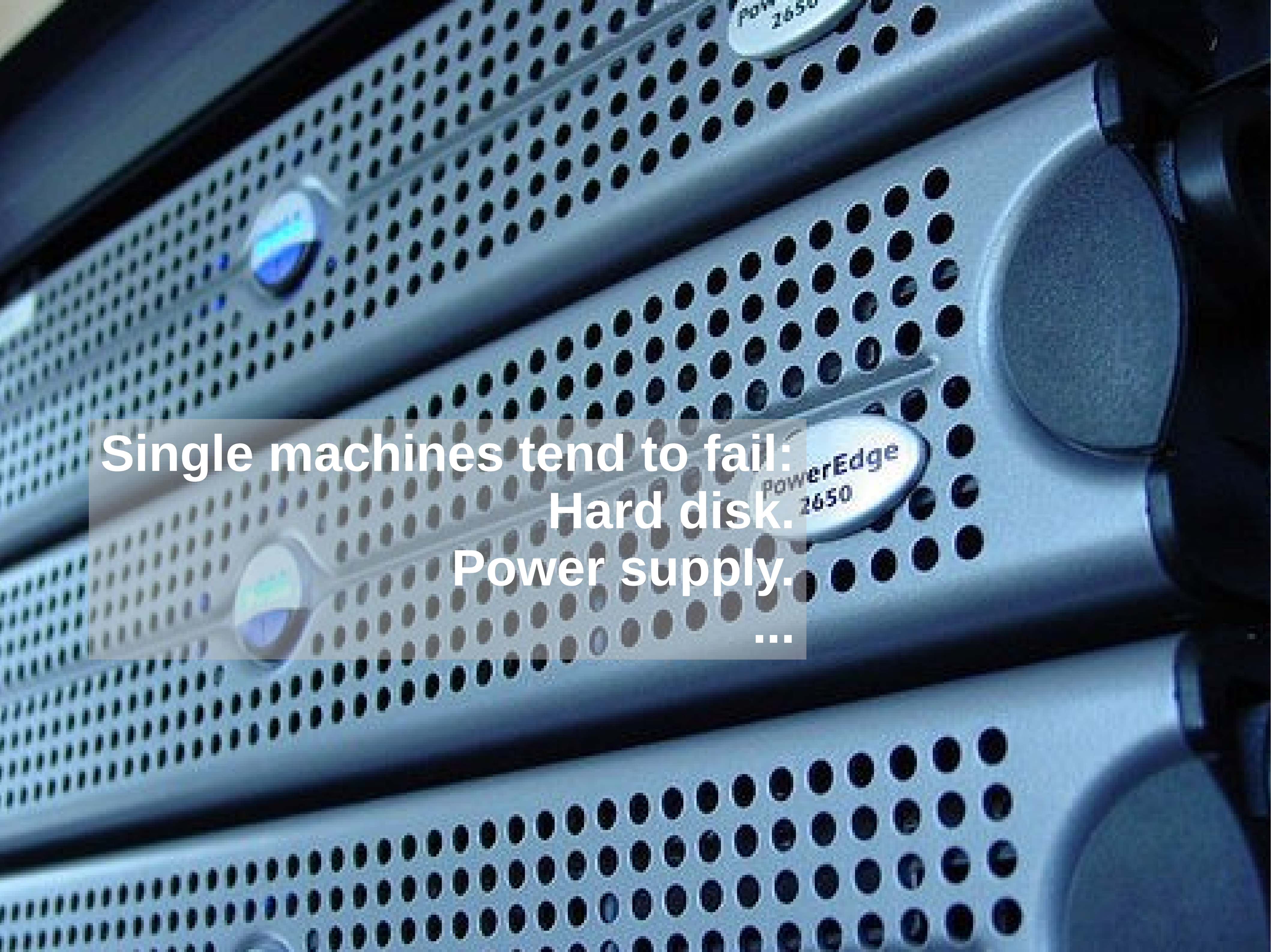
January 8, 2008 by Pink Sherbet Photography
http://www.flickr.com/photos/pinksherbet/2177961471/

# Massive data as in:

Cannot be stored on single machine.
Takes too long to process in serial.

Idea: Use multiple machines.

# Challenges when scaling out.



Single machines tend to fail:
Hard disk.
Power supply.
...

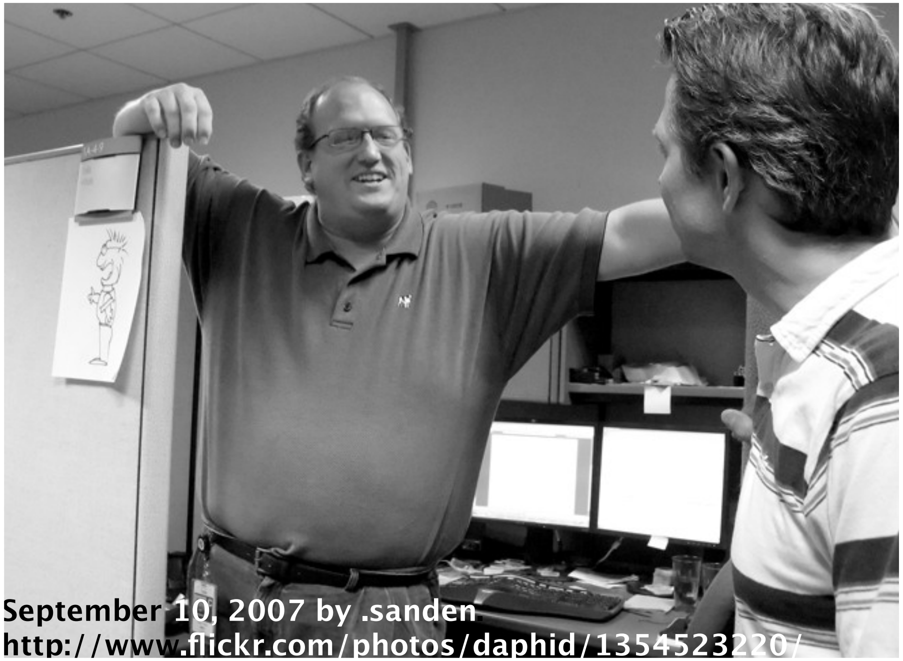
# More machines – increased failure probability.

January 11, 2007, skreuzer
http://www.flickr.com/photos/skreuzer/354316053/

# Requirements

- Built-in backup.
- Built-in failover.

# Typical developer



September 10, 2007 by .sanden
http://www.flickr.com/photos/daphid/1354523220/

- Has never dealt with large (petabytes) amount of data.
- Has no thorough understanding of parallel programming.
- Has no time to make software production ready.

# Requirements

- Built-in backup.
- Built-in failover.
- Easy to use.
- Parallel on rails.

February 29, 2008 by Thomas Claveirole
http://www.flickr.com/photos/thomasclaveirole/2300932656/

http://www.flickr.com/photos/jaaronfarr/3384940437/
March 25, 2009 by jaaron

http://www.flickr.com/photos
/jaaronfarr/3385756482/
March 25, 2009 by jaaron

May 1, 2007 by danny angus
http://www.flickr.com/photos/killerbees/479864437/

# Requirements

- Built-in backup.
- Built-in failover.

- Easy to use.
- Parallel on rails.

- Active development.

http://www.flickr.com/photos/cspowers/282944734/ by cspowers on October 29, 2006

# Requirements

- Built-in backup.
- Built-in failover.

- Easy to administrate.
- Single system.

- Easy to use.
- Parallel on rails.

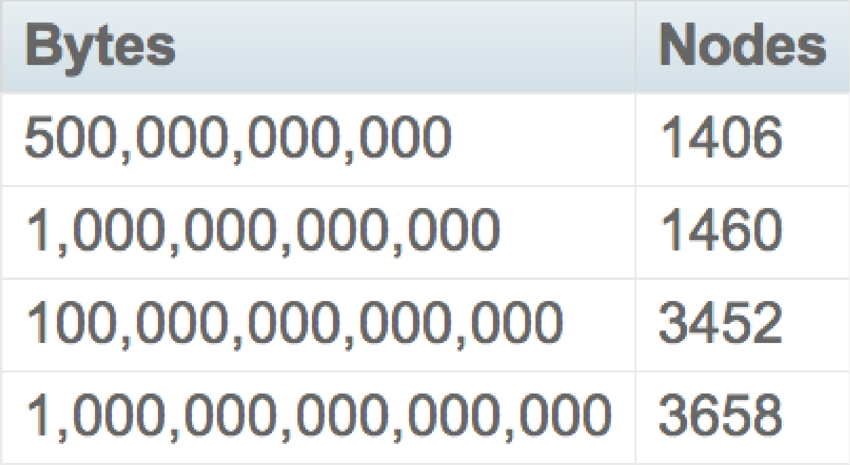
- Active development.

Easy distributed programming.

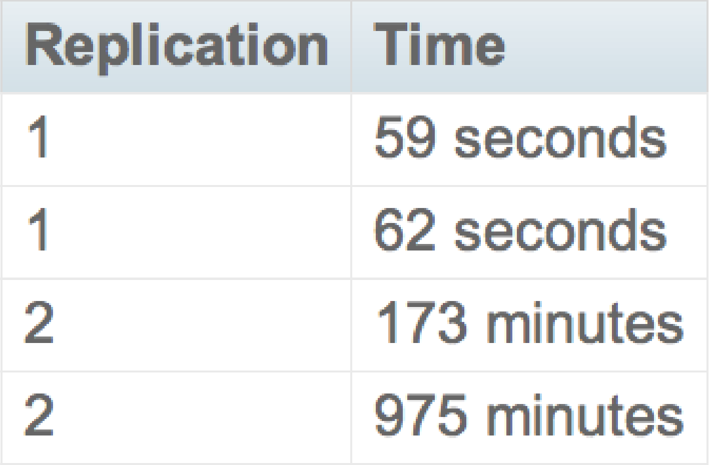
Well known in industry and research.

Scales well beyond 1000 nodes.

# Petabyte sorting benchmark



| Bytes | Nodes |
|---|---|
| 500,000,000,000 | 1406 |
| 1,000,000,000,000 | 1460 |
| 100,000,000,000,000 | 3452 |
| 1,000,000,000,000,000 | 3658 |



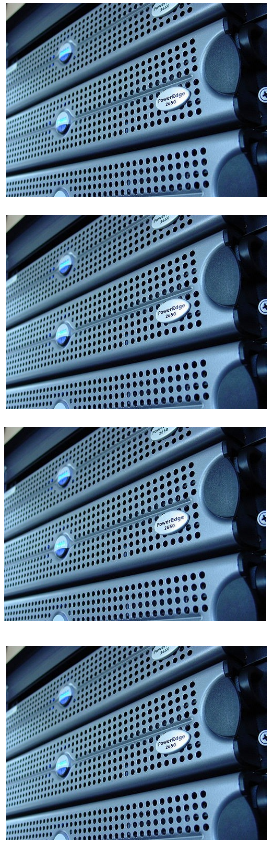
| Replication | Time |
|---|---|
| 1 | 59 seconds |
| 1 | 62 seconds |
| 2 | 173 minutes |
| 2 | 975 minutes |

Per node: 2 quad core Xeons @ 2.5ghz, 4 SATA disks, 8G RAM (upgraded to 16GB before petabyte sort), 1 gigabit ethernet.

Per Rack: 40 nodes, 8 gigabit ethernet uplinks.

# Assumptions:

Data to process does not fit on one node.
Each node is commodity hardware.
Failure happens.

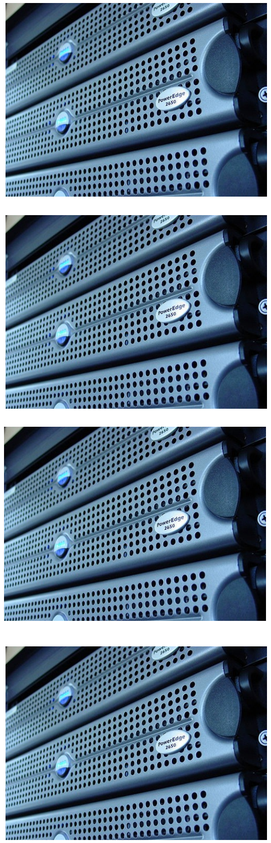


# Ideas:

Distribute filesystem.
Built in replication.
Automatic failover in case of failure.

# Assumptions:

Moving data is expensive.
Moving computation is cheap.
Distributed computation is easy.

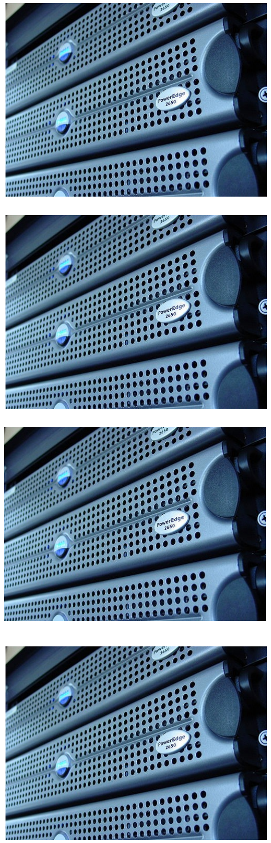


# Ideas:

Move computation to data.
Write software that is easy to distribute.

# Assumptions:

Systems run on spinning hard disks.
Disk seek >> disk scan.



# Ideas:

Improve support for large files.
File system API makes scanning easy.

# Data storage options

- Structured, relational .
  - Customer data.
  - Bug database.

- Semi-structured data:
  - Documents.
  - Independent rows.

- Continuous files.
  - Log data.
  - Document Stream.

# Store in RDBMS?

- Possible.
- Becomes expensive pretty quickly.

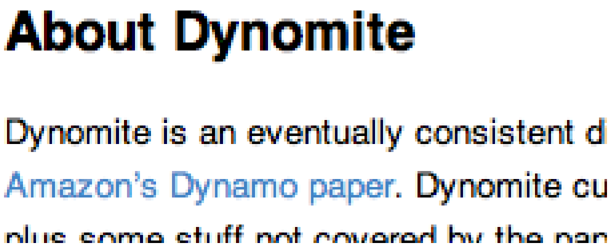
# Store in Hadoop DFS?

- Optimised for LARGE files.
- Throughput vs. latency.

# Something in between?

- Transactions – can we do without?

- Joins – some applications don't need them.



## About Dynomite

Dynomite is an eventually consistent d
Amazon's Dynamo paper. Dynomite cu
plus some stuff not covered by the pap

# From data to information.

- ✓ Start collecting and storing your data.
- Analyse and understand your data.
- Answer more complex questions.

# Understanding your data

- Data profiling.

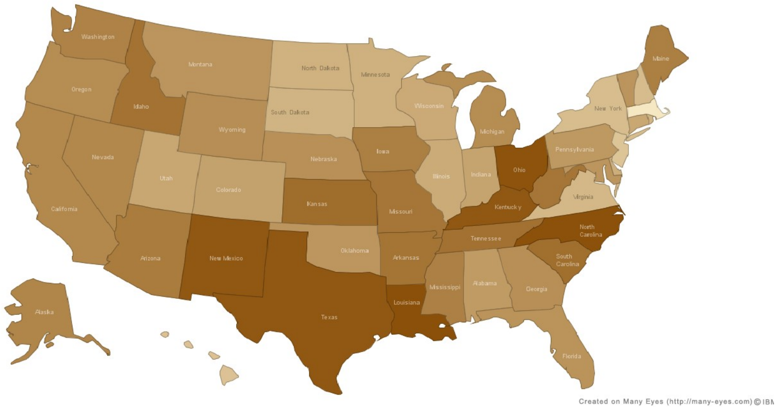
- Goals:
  - Identify usual behaviour.
  - Find exceptional cases.

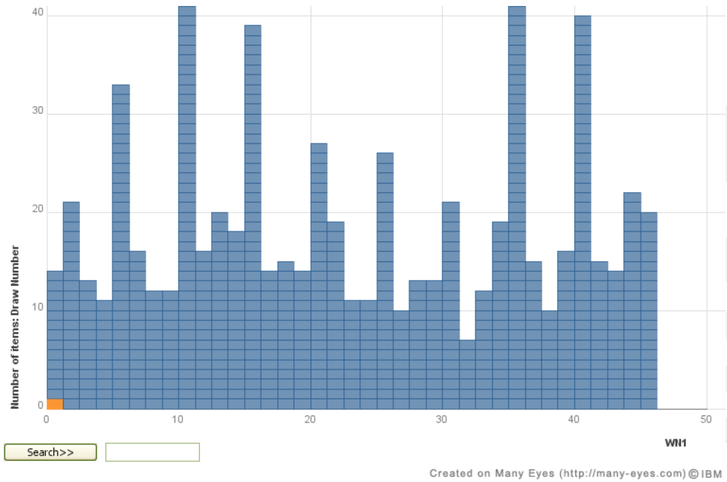
- Exact questions depend on domain.

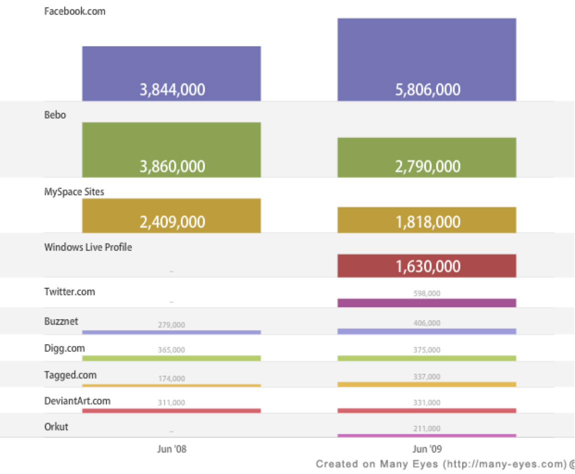
# Example questions

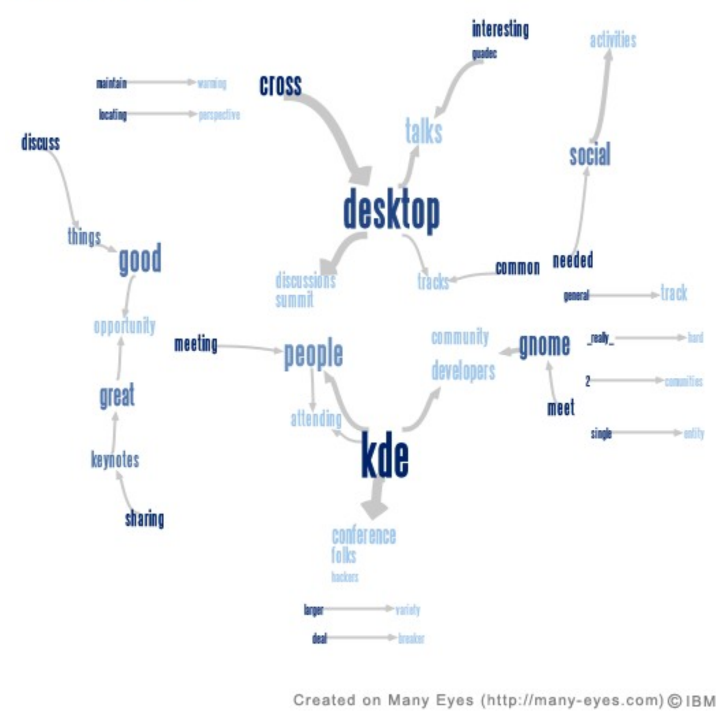
- Structured data:
  - Shopping: Amount of money usually spent.
  - Average age of your customers.
  - Min/Max number of shopping sessions.
- Textual documents:
  - Average length of documents.
  - Distribution of document topics.
  - Distribution of authors.

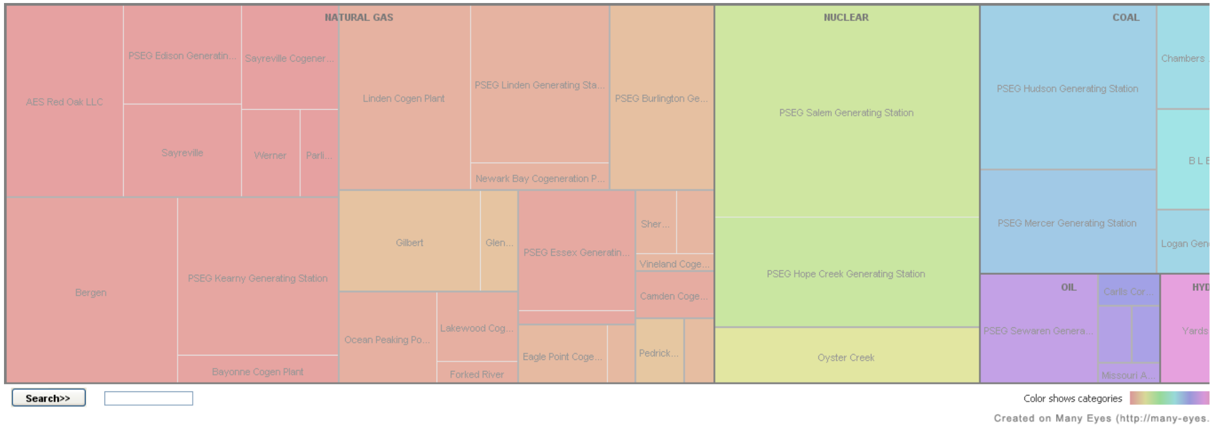
# Visualizations help







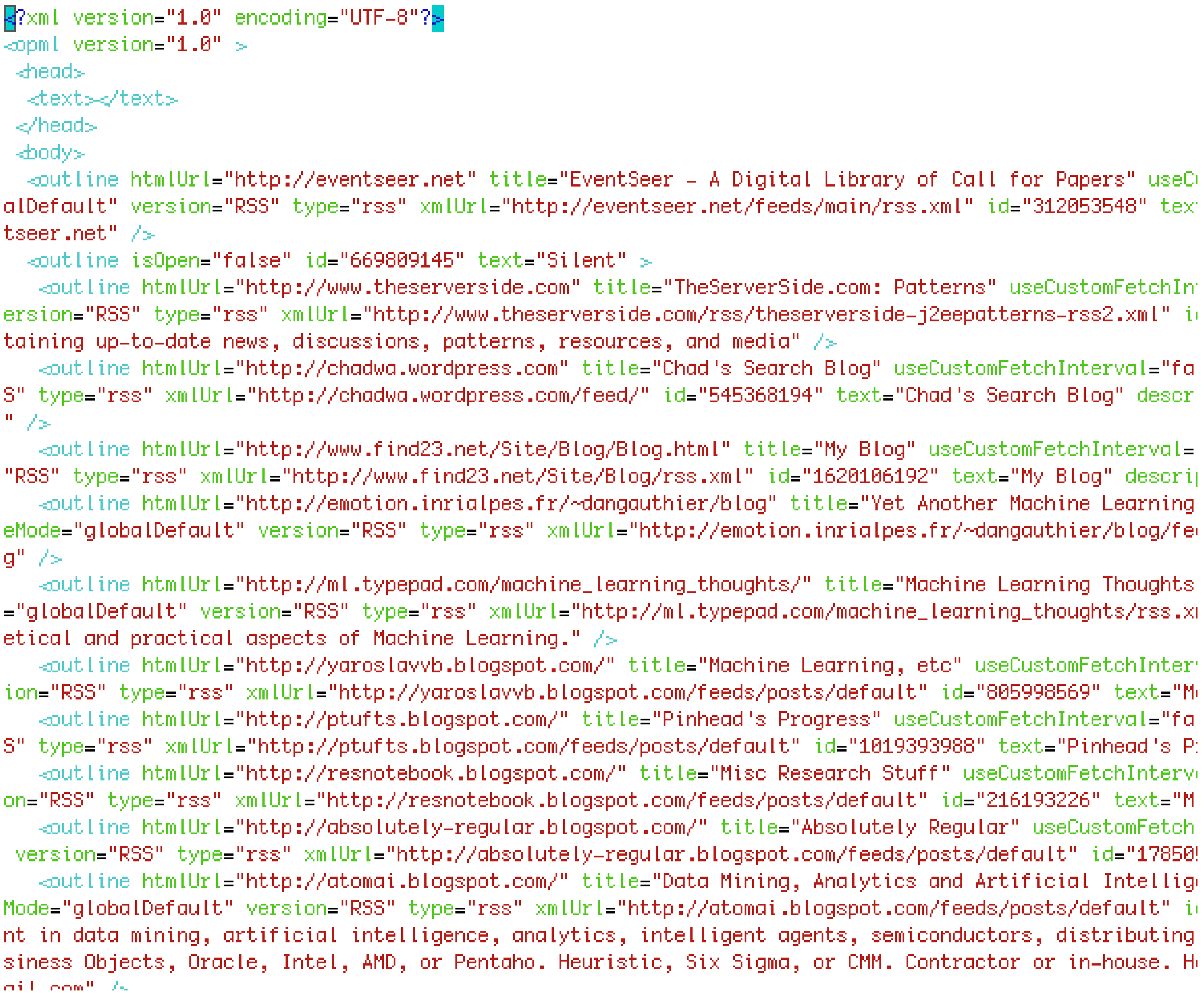




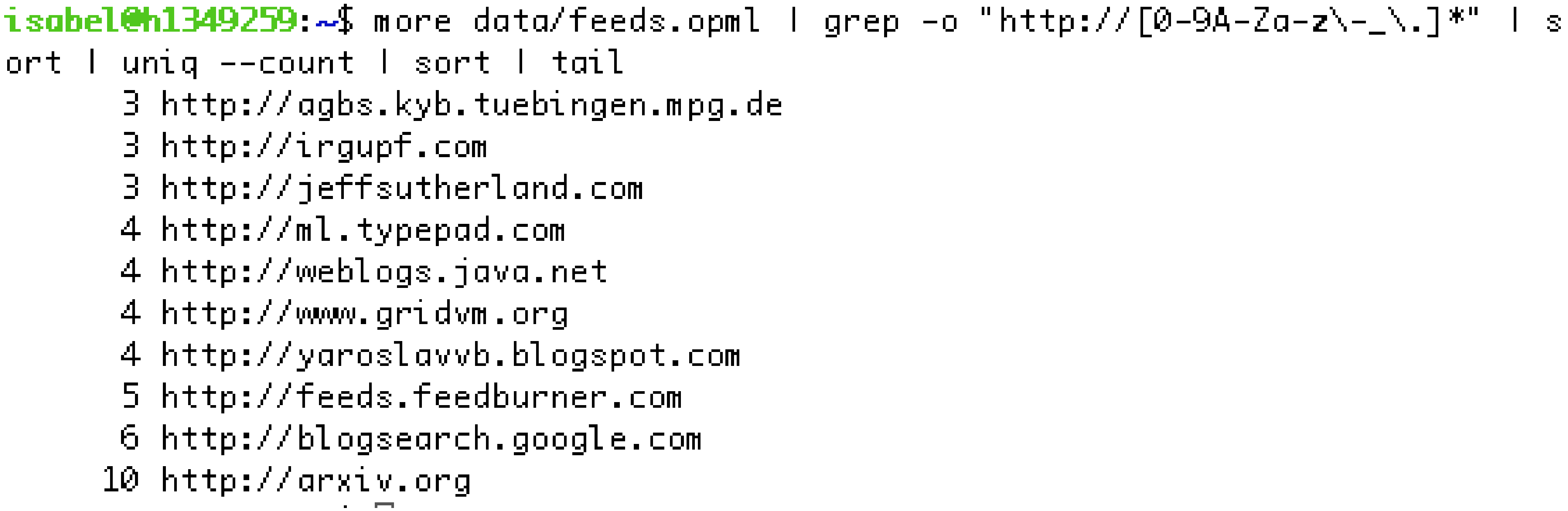
# Understanding your data

- Structured data in RDBMS:
  - Functionality built-in (min, max etc.)
- Unstructured or Semistructured data in HDFS:
  - Write analysis code in Java. (Map/Reduce jobs).
  - Use higher level language.

# Map/Reduce by example



```
<?xml version="1.0" encoding="UTF-8"?>
<opml version="1.0" >
 <head>
  <text></text>
 </head>
 <body>
  <outline htmlUrl="http://eventseer.net" title="EventSeer - A Digital Library of Call for Papers" useC
alDefault" version="RSS" type="rss" xmlUrl="http://eventseer.net/feeds/main/rss.xml" id="312053548" tex
tseer.net" />
  <outline isOpen="false" id="669809145" text="Silent" >
   <outline htmlUrl="http://www.theserverside.com" title="TheServerSide.com: Patterns" useCustomFetchIn
ersion="RSS" type="rss" xmlUrl="http://www.theserverside.com/rss/theserverside-j2eepatterns-rss2.xml" i
taining up-to-date news, discussions, patterns, resources, and media" />
   <outline htmlUrl="http://chadwa.wordpress.com" title="Chad's Search Blog" useCustomFetchInterval="fa
S" type="rss" xmlUrl="http://chadwa.wordpress.com/feed/" id="545368194" text="Chad's Search Blog" descr
" />
   <outline htmlUrl="http://www.find23.net/Site/Blog/Blog.html" title="My Blog" useCustomFetchInterval=
"RSS" type="rss" xmlUrl="http://www.find23.net/Site/Blog/rss.xml" id="1620106192" text="My Blog" descri
   <outline htmlUrl="http://emotion.inrialpes.fr/~dangauthier/blog" title="Yet Another Machine Learning
eMode="globalDefault" version="RSS" type="rss" xmlUrl="http://emotion.inrialpes.fr/~dangauthier/blog/fe
g" />
   <outline htmlUrl="http://ml.typepad.com/machine_learning_thoughts/" title="Machine Learning Thoughts
="globalDefault" version="RSS" type="rss" xmlUrl="http://ml.typepad.com/machine_learning_thoughts/rss.x
etical and practical aspects of Machine Learning." />
   <outline htmlUrl="http://yaroslavvb.blogspot.com/" title="Machine Learning, etc" useCustomFetchInter
ion="RSS" type="rss" xmlUrl="http://yaroslavvb.blogspot.com/feeds/posts/default" id="805998569" text="M
   <outline htmlUrl="http://ptufts.blogspot.com/" title="Pinhead's Progress" useCustomFetchInterval="fa
S" type="rss" xmlUrl="http://ptufts.blogspot.com/feeds/posts/default" id="1019393988" text="Pinhead's P
   <outline htmlUrl="http://resnotebook.blogspot.com/" title="Misc Research Stuff" useCustomFetchInterv
on="RSS" type="rss" xmlUrl="http://resnotebook.blogspot.com/feeds/posts/default" id="216193226" text="M
   <outline htmlUrl="http://absolutely-regular.blogspot.com/" title="Absolutely Regular" useCustomFetch
 version="RSS" type="rss" xmlUrl="http://absolutely-regular.blogspot.com/feeds/posts/default" id="17850
   <outline htmlUrl="http://atomai.blogspot.com/" title="Data Mining, Analytics and Artificial Intellig
Mode="globalDefault" version="RSS" type="rss" xmlUrl="http://atomai.blogspot.com/feeds/posts/default" i
nt in data mining, artificial intelligence, analytics, intelligent agents, semiconductors, distributing
siness Objects, Oracle, Intel, AMD, or Pentaho. Heuristic, Six Sigma, or CMM. Contractor or in-house. H
ail.com" />
```



```
isabel@h1349259:~$ more data/feeds.opml | grep -o "http://[0-9A-Za-z\-_\.]*" | s
ort | uniq --count | sort | tail
      3 http://agbs.kyb.tuebingen.mpg.de
      3 http://irgupf.com
      3 http://jeffsutherland.com
      4 http://ml.typepad.com
      4 http://weblogs.java.net
      4 http://www.gridvm.org
      4 http://yaroslavvb.blogspot.com
      5 http://feeds.feedburner.com
      6 http://blogsearch.google.com
     10 http://arxiv.org
```

```
pattern=”http://[0-9A-Za-z\-_\.]*”

grep -o "$pattern" feeds.opml     | sort          | uniq --count
```

```
pattern=”http://[0-9A-Za-z\-_\.]*”
```

```
grep -o "$pattern" feeds.opml
```

MAP

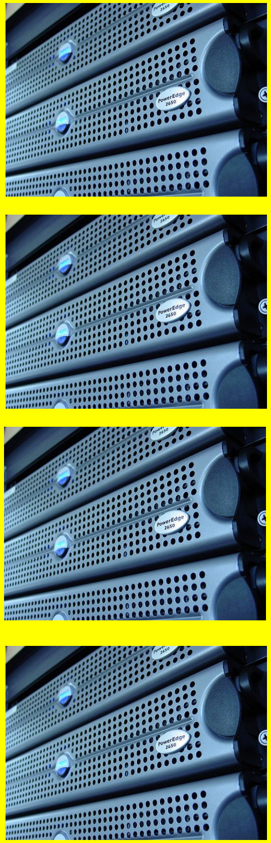
```
| sort
```

| SHUFFLE

```
| uniq --count
```

| REDUCE

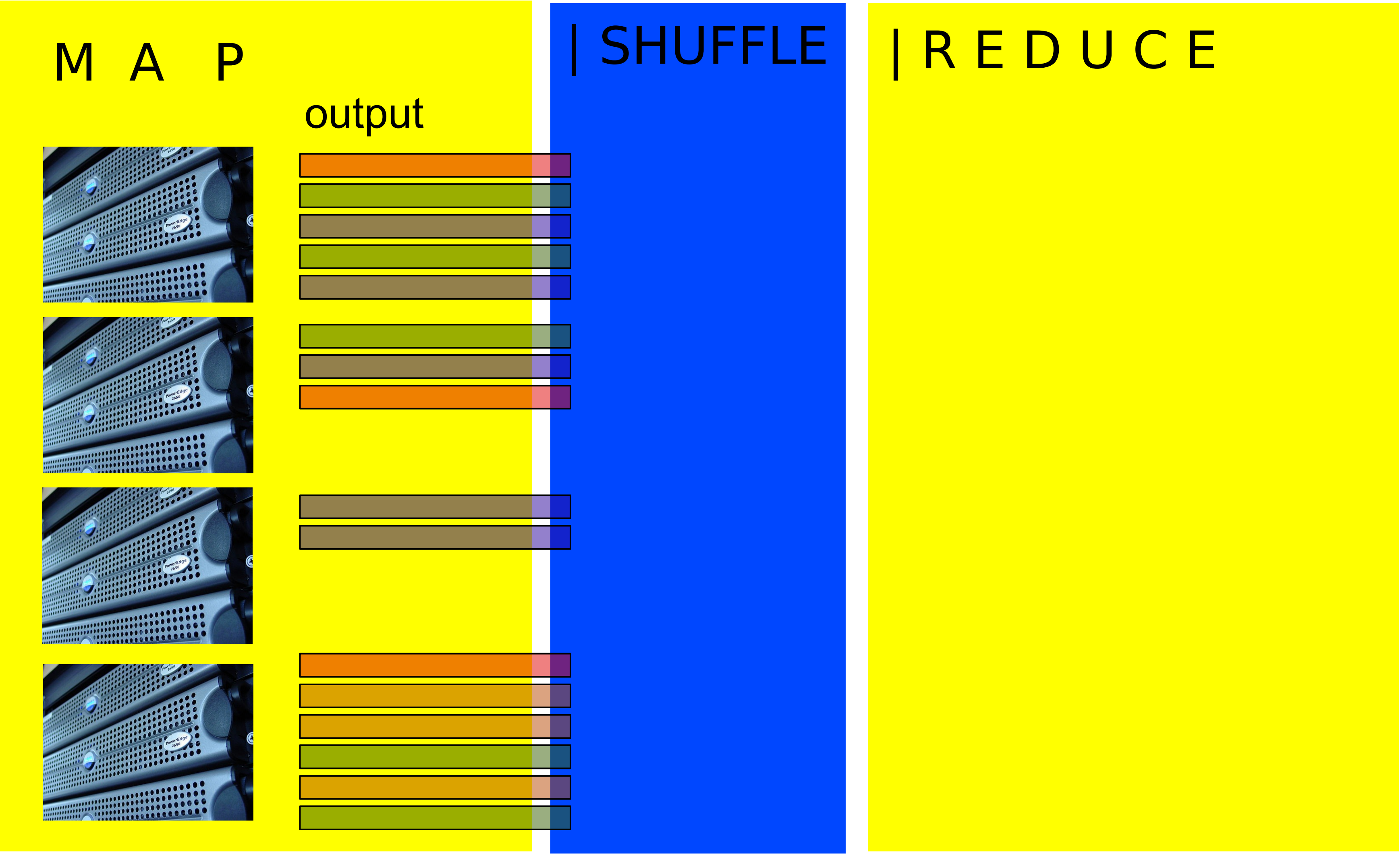
MAP



| SHUFFLE

| REDUCE

Local to data.



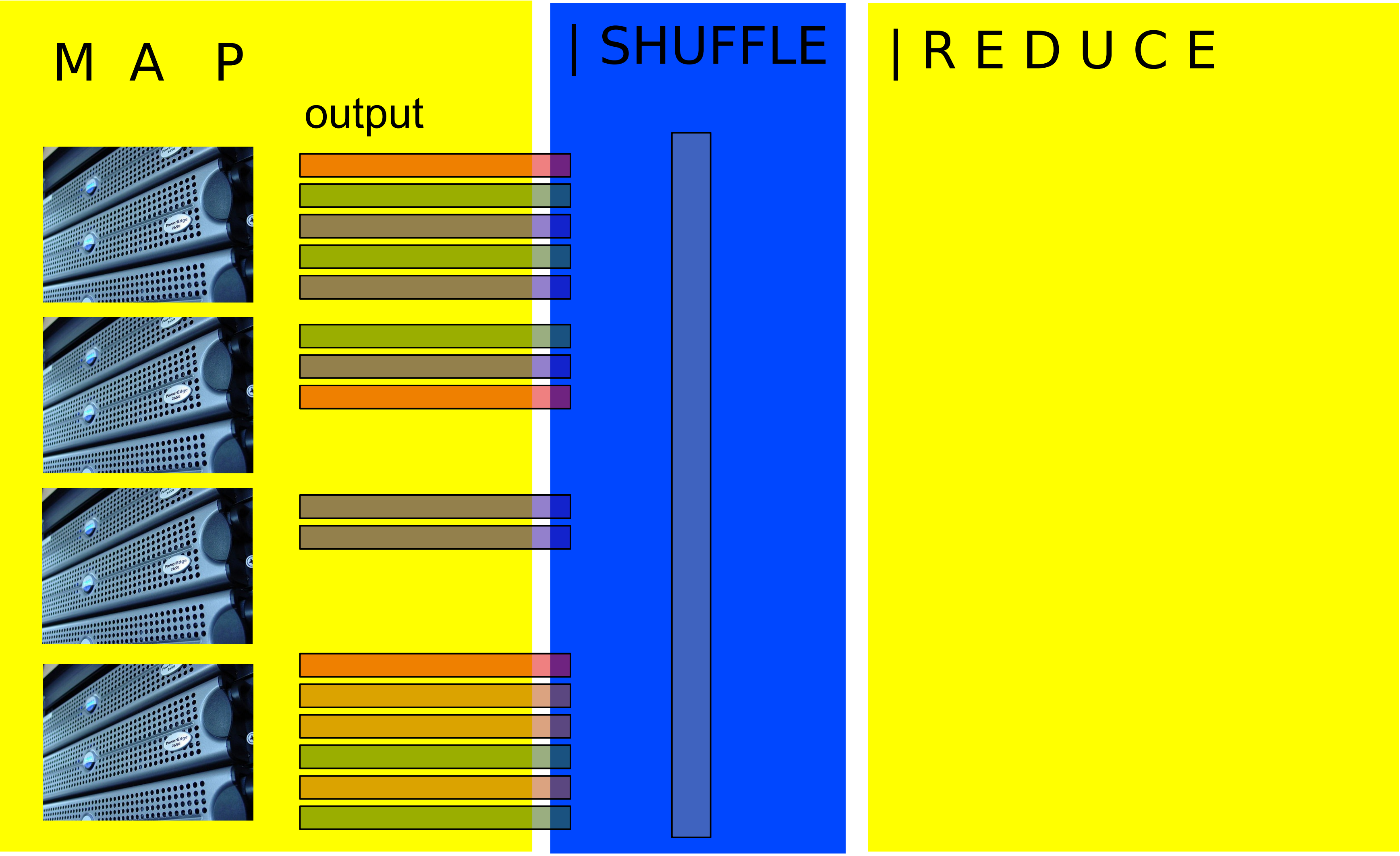
MAP

output

| SHUFFLE

| REDUCE

Local to data.
Outputs a lot less data.
Output can cheaply move.



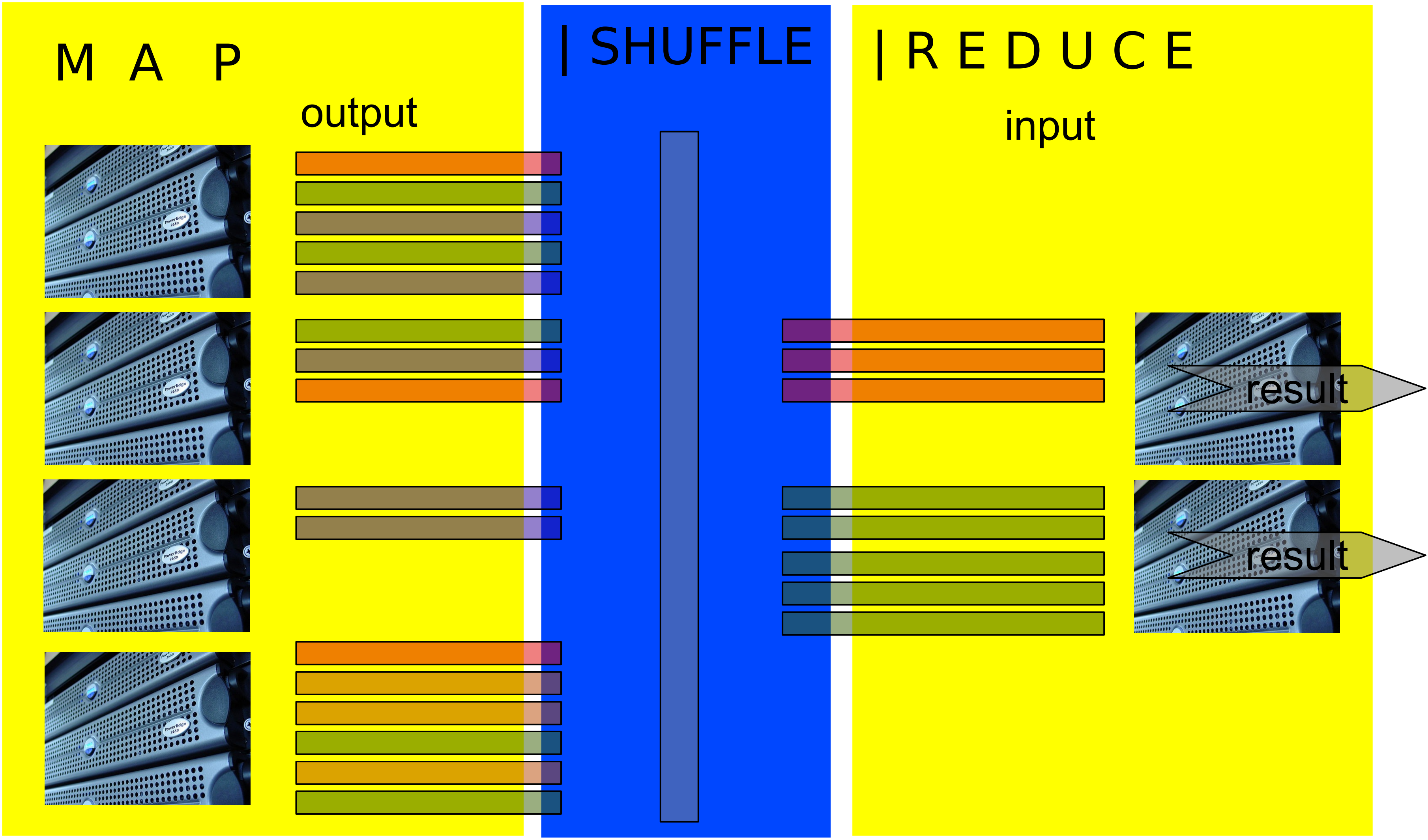
MAP

output

| SHUFFLE

| REDUCE

Local to data.
Outputs a lot less data.
Output can cheaply move.



MAP

output

| SHUFFLE

| REDUCE

input

Local to data.
Outputs a lot less data.
Output can cheaply move.

Shuffle sorts input by key.
Reduces output significantly.

```
private IntWritable one = new IntWritable(1);
private Text hostname = new Text();
```

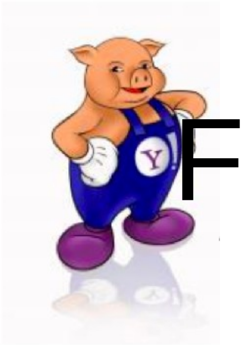
```
public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> output,
Reporter reporter) throws IOException {
  String line = value.toString();
  StringTokenizer tokenizer = new StringTokenizer(line);
  while (tokenizer.hasMoreTokens()) {
    hostname.set(getHostname(tokenizer.nextToken()));
    output.collect(hostname, one);
  }
}
```

```
public void reduce(Text key, Iterator<IntWritable>
values, OutputCollector<Text, IntWritable> output,
Reporter reporter) throws IOException {
  int sum = 0;
  while (values.hasNext()) {
    sum += values.next().get();
  }
  output.collect(key, new IntWritable(sum));
}
```

# Higher level languages.

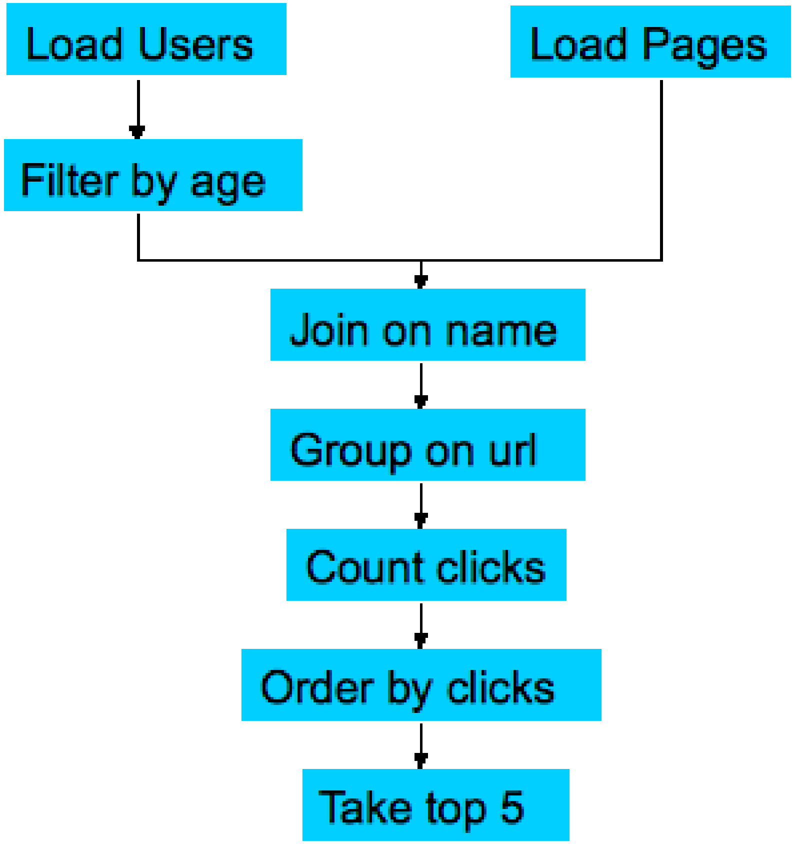
Cascading

Y!



# Filtering/ Aggregating in Hadoop

Suppose you have user data in one file, website data in another, and you need to find the top 5 most visited pages by users aged 18 - 25.



Load Users

Load Pages

Filter by age

Join on name

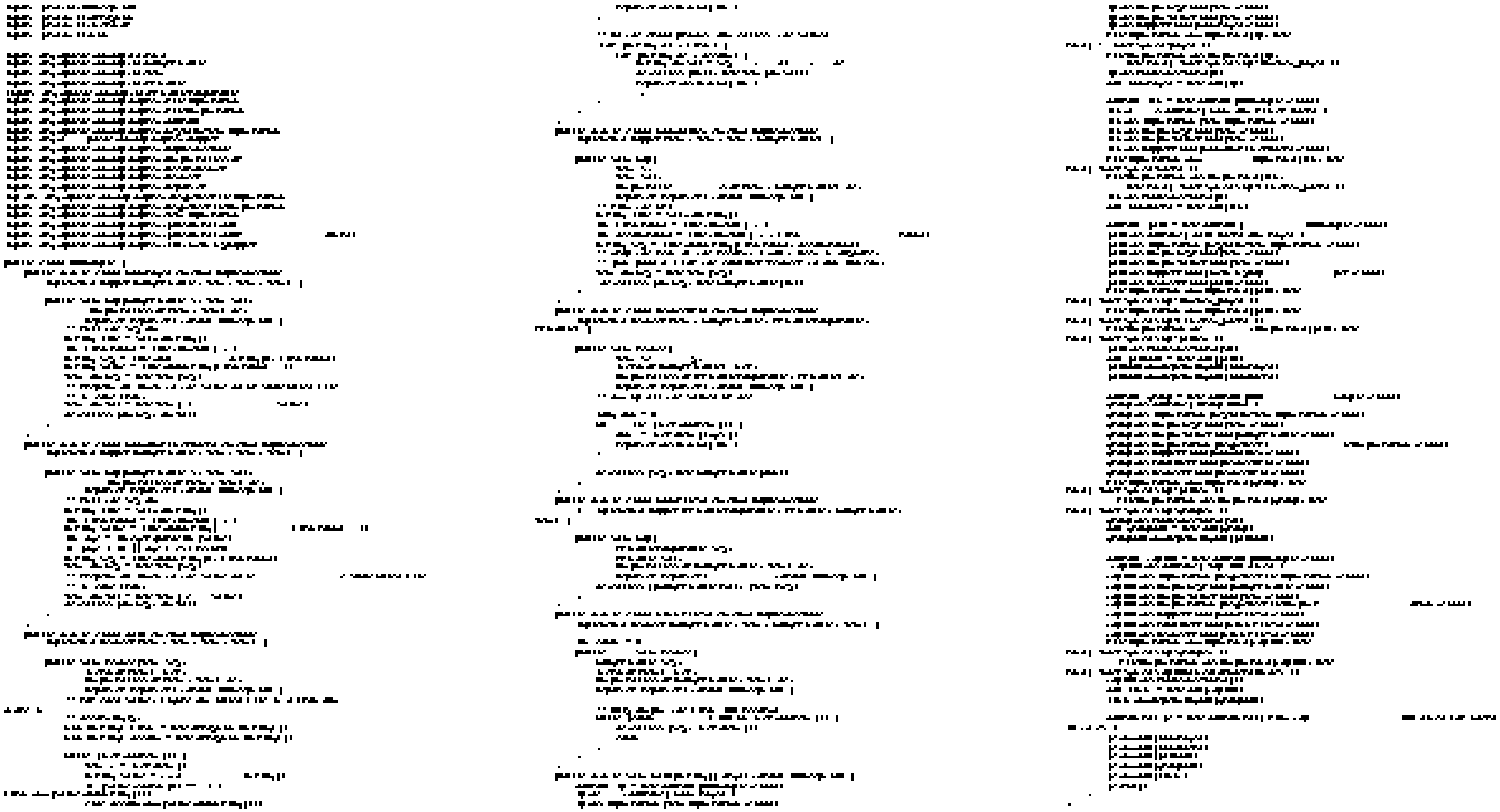
Group on url

Count clicks

Order by clicks

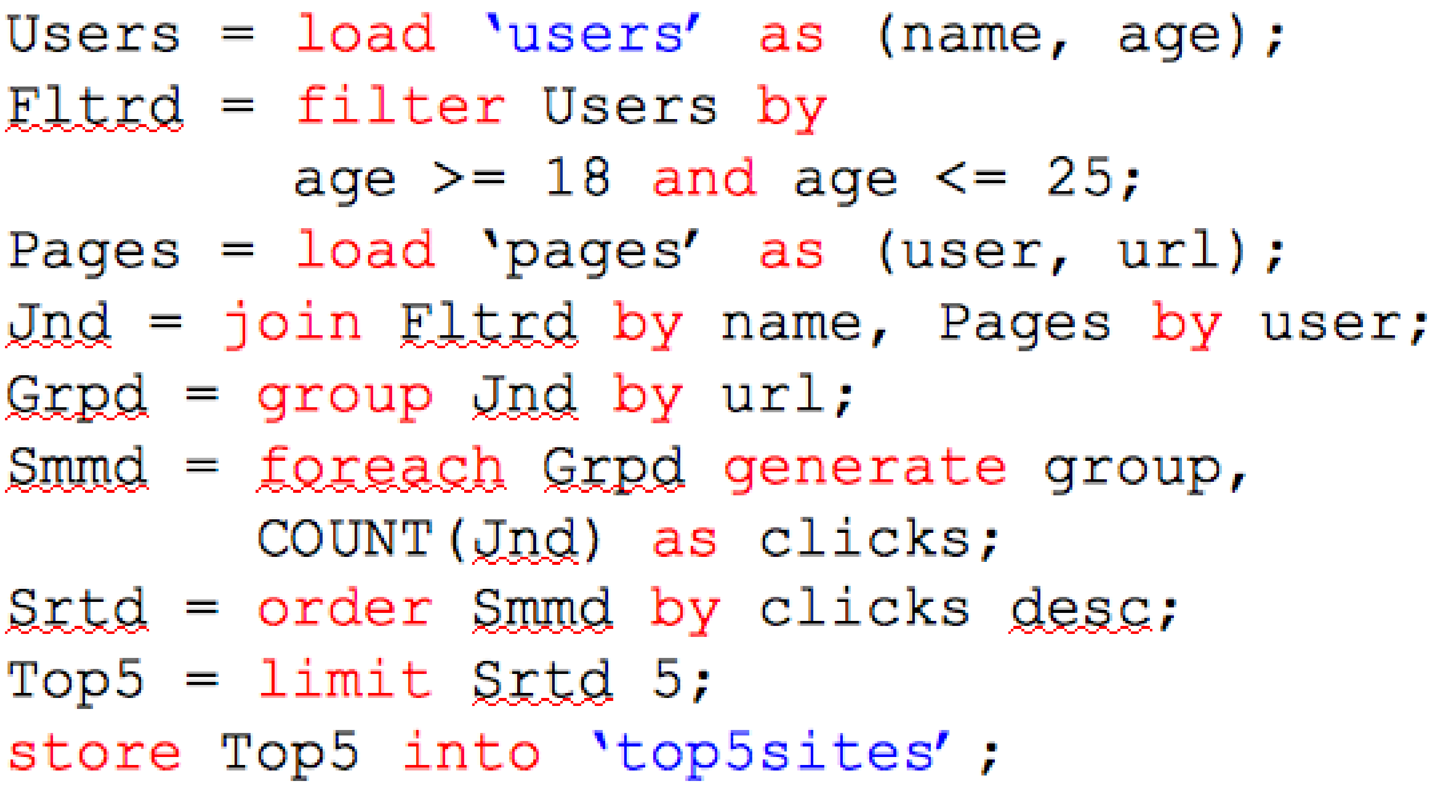
Take top 5

Example from PIG presentation at Apache Con EU 2009



[illegible]

Example from PIG presentation at Apache Con EU 2009



```
Users = load ‘users’ as (name, age);
Fltrd = filter Users by
        age >= 18 and age <= 25;
Pages = load ‘pages’ as (user, url);
Jnd = join Fltrd by name, Pages by user;
Grpd = group Jnd by url;
Smmd = foreach Grpd generate group,
       COUNT(Jnd) as clicks;
Srtd = order Smmd by clicks desc;
Top5 = limit Srtd 5;
store Top5 into ‘top5sites’;
```

Example from PIG presentation at Apache Con EU 2009

# From data to information.

- ✓ Start collecting and storing your data.
- ✓ Analyse and understand your data.
- Answer more complex questions.

# More complex questions

- Which products are commonly bought together.
- What groups of search results were returned.
- Predict probability of user clicking an ad.
- Identify emerging topics in news stories.
- Find source code commonly changed together.
- Identify malicious access patterns to servers.

# Machine learning – what's that?

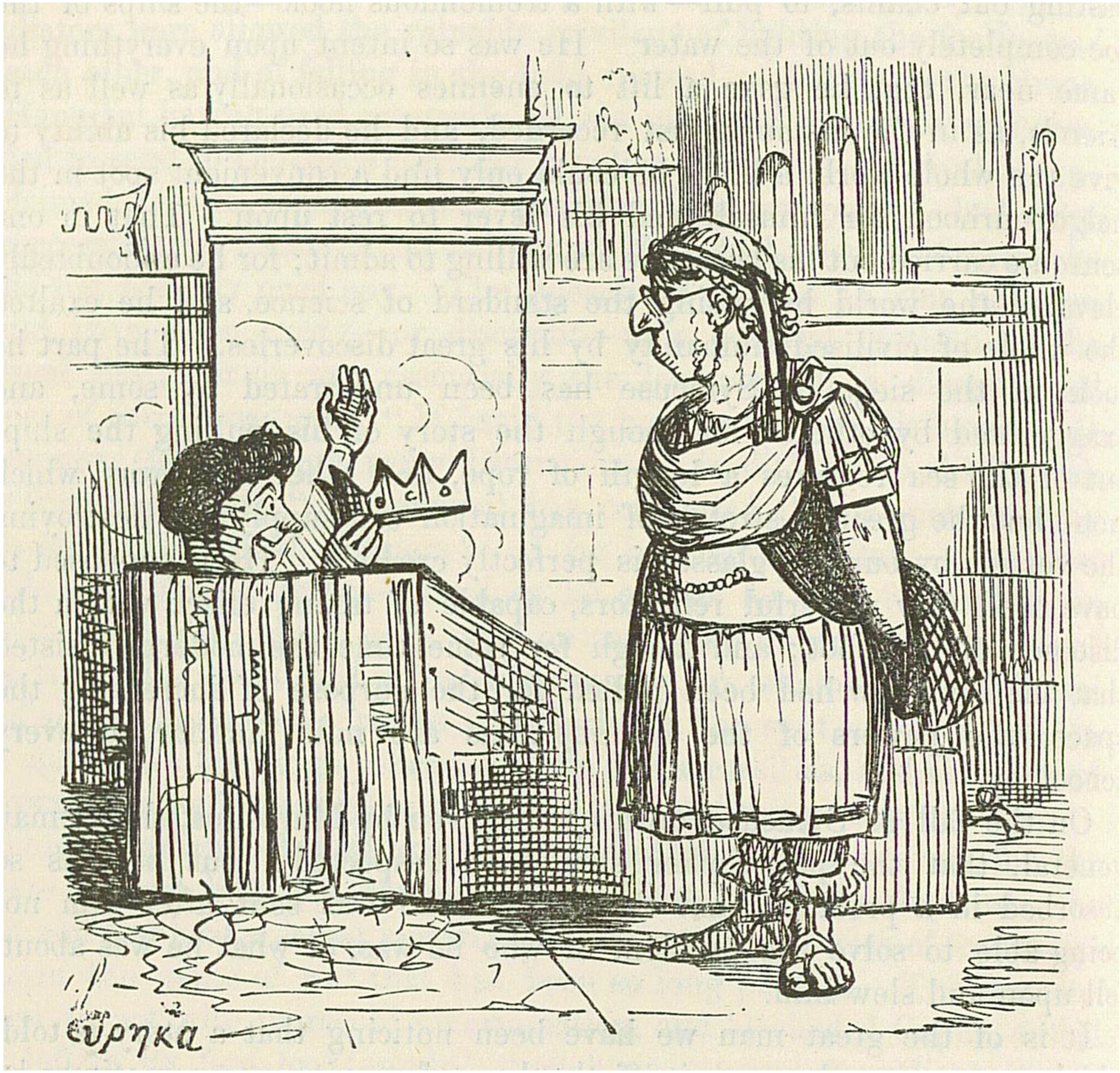
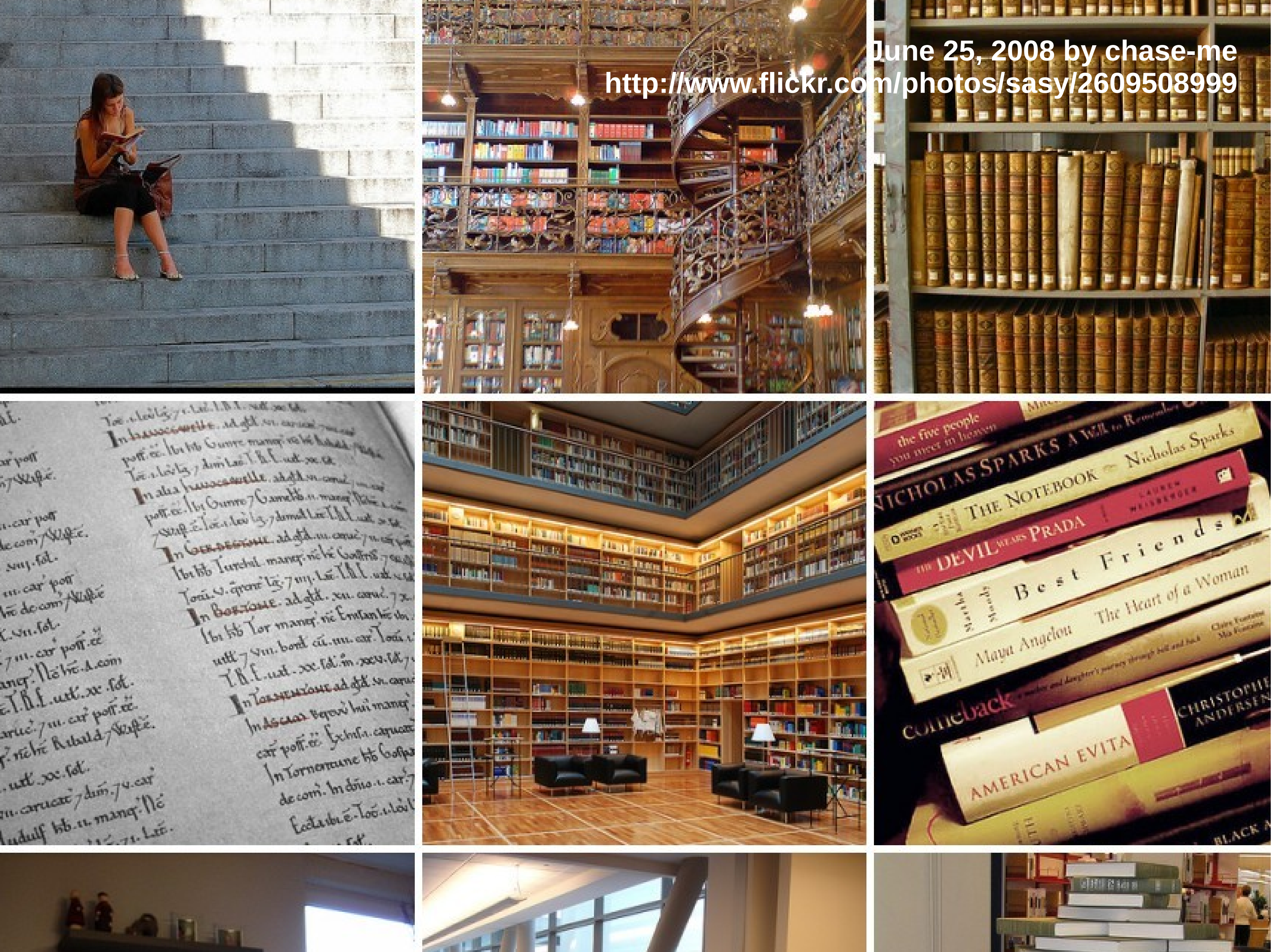


**Image by John Leech, from: The Comic History of Rome by Gilbert Abbott A Beckett.**
**Bradbury, Evans & Co, London, 1850s**
**Archimedes taking a Warm Bath**

# Archimedes model of nature

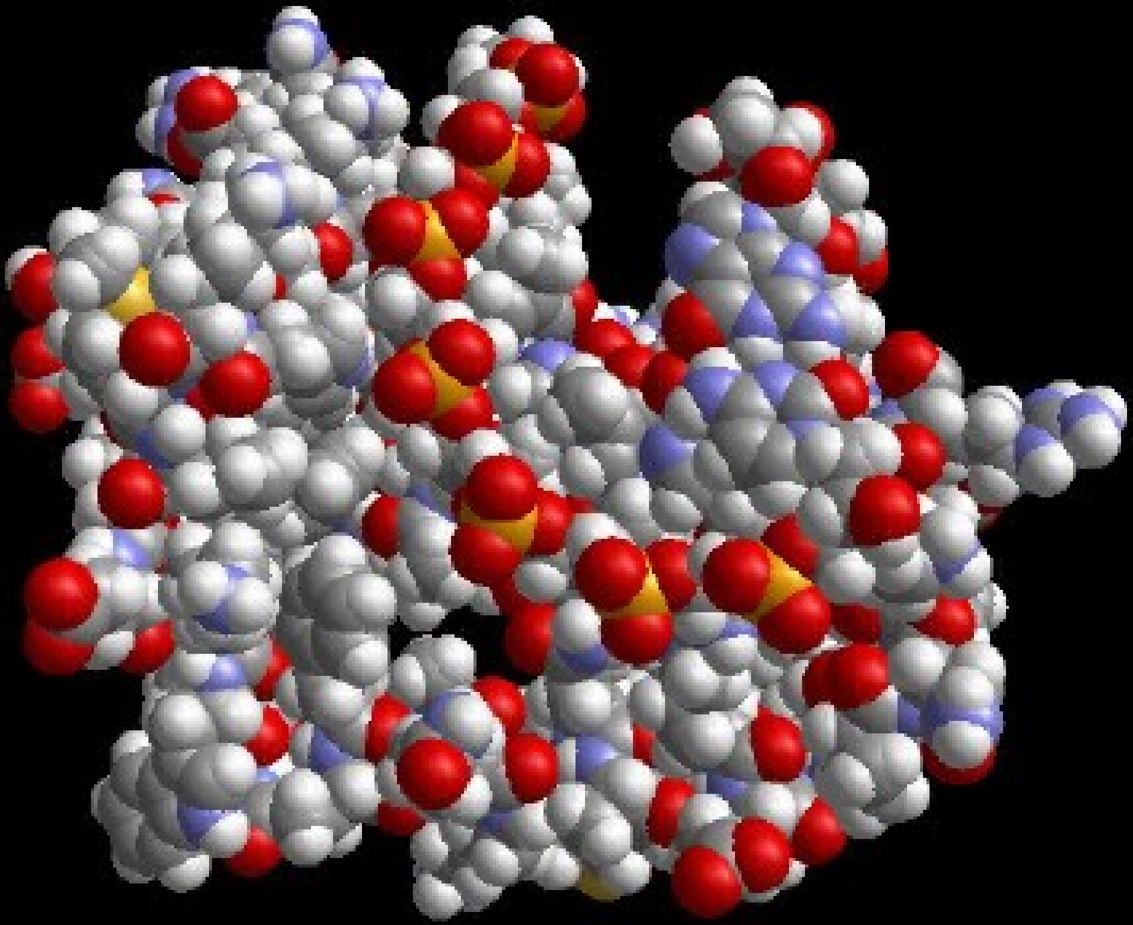
$$\frac{\mathit{Density\ of\ Object}}{\mathit{Density\ of\ Fluid}} = .$$

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

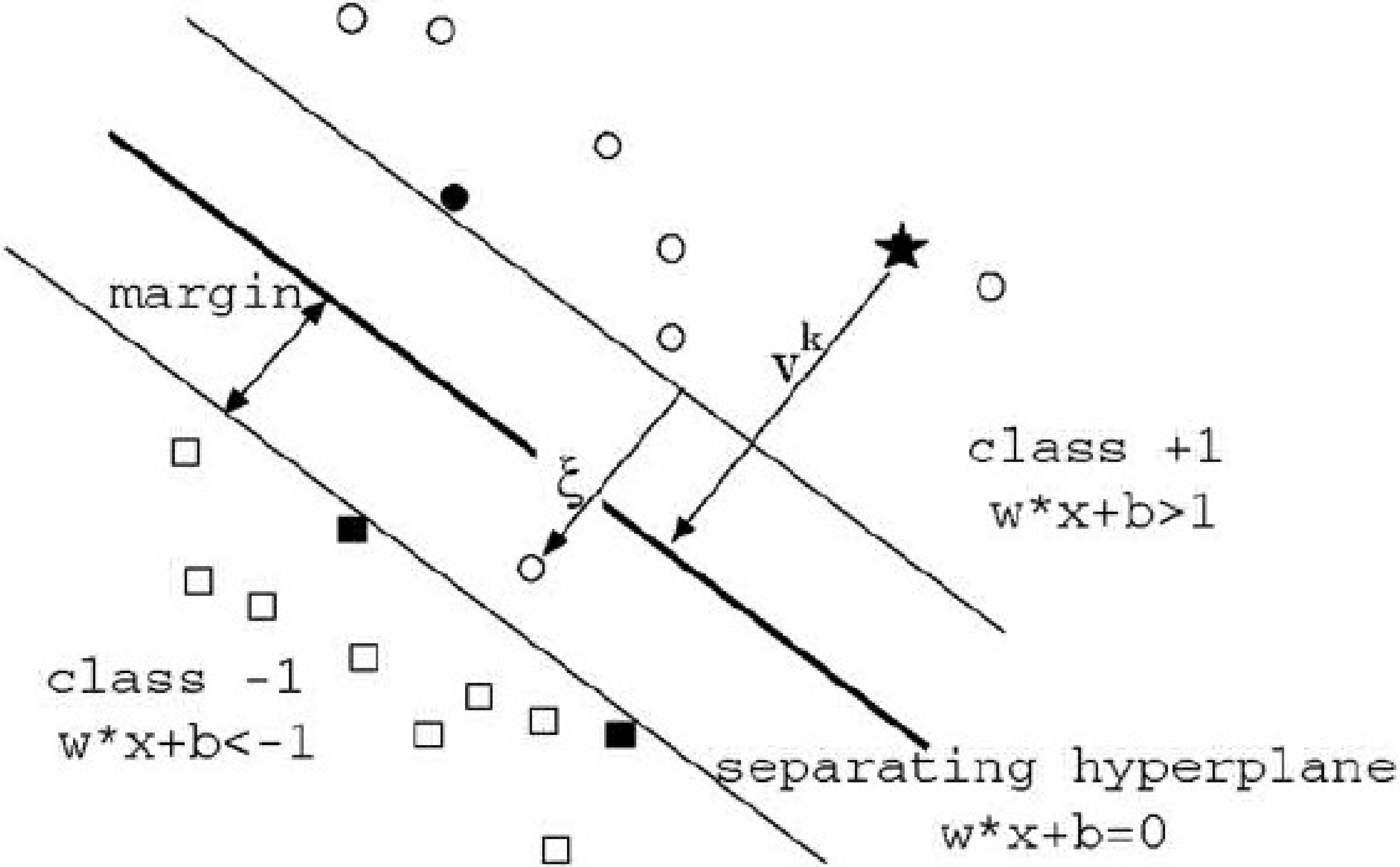


June 25, 2008 by chase-me
http://www.flickr.com/photos/sasy/2609508999

March 28, 2007 by dullhunk
http://www.flickr.com/photos/dullhunk/437551254



# An SVM's model of nature



# Scaling machine learning.

# Contributions need not be Java based:

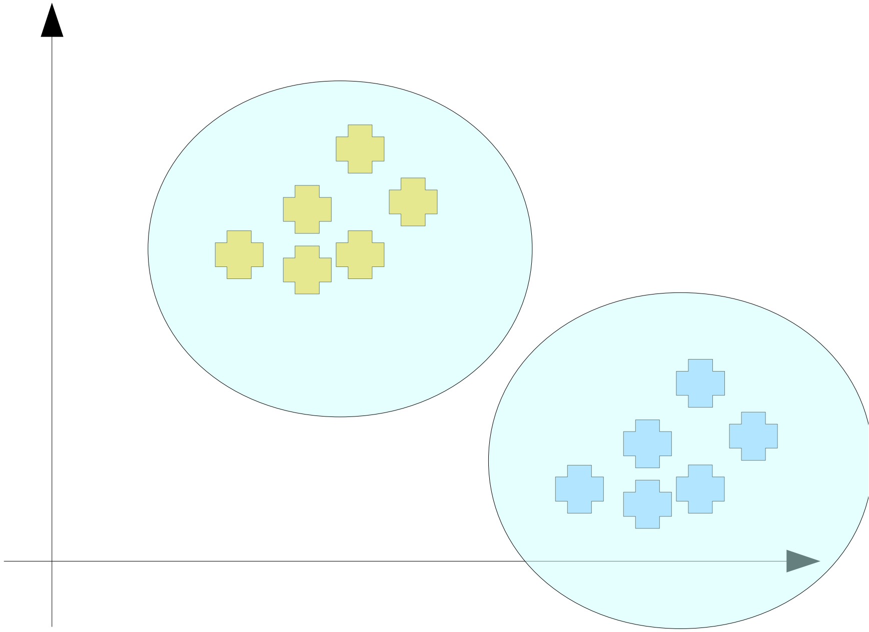
PIG, JAQL, Cascading, ...?

- Industry ready.
  - Friendly license.
  - Scalable.
- Easy to use.
  - Well documented.
  - Well maintained by healthy and active community.
- Easy to extend and contribute to.
  - Automated tests.
  - Easy to build and deploy.

# What does Mahout have to offer.

# Discover groups of items

- Group items by similarity.



- Examples:
  - Group news articles by topic.
  - Find developers with similar interests.
  - Discovery of groups of related search results.

# Discover groups of similar items

- Canopy.
- k-Means.
- Fuzzy k-Means.
- Dirichlet based.
- Others upcoming.

# Identify dominant topics

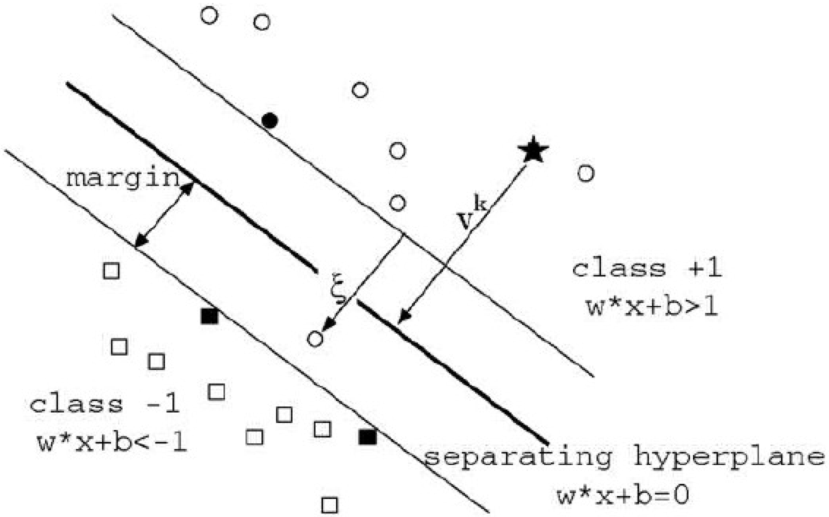
- Given a dataset of texts, identify main topics.

Algorithms: Parallel LDA

- Examples:
  - Dominant topics in set of mails.
  - Identify news message categories.

# Assign items to defined categories.

- Given pre-defined categories, assin items to it.



- Examples:
  - Spam mail classification.
  - Discovery of images depicting humans.

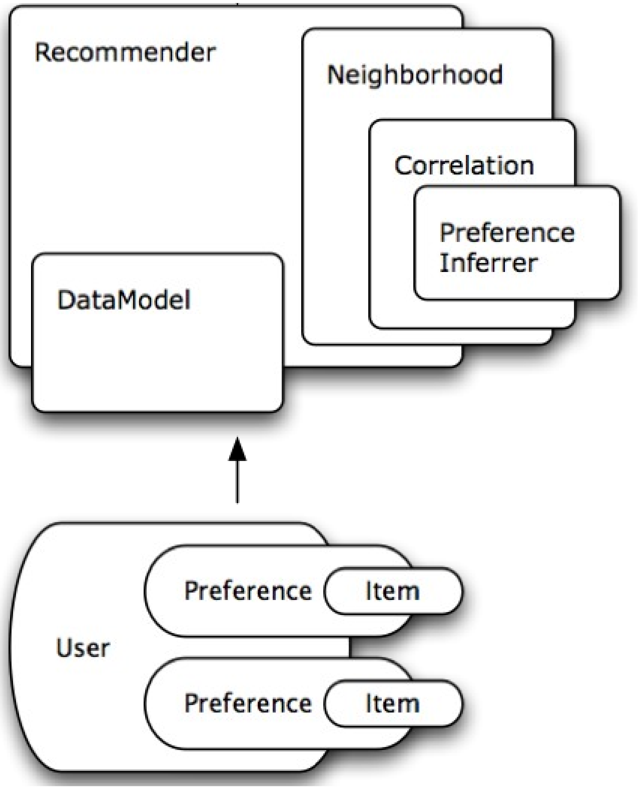
# Assign items to defined categories.

- Naïve Bayes.
- Complementary naïve bayes.

- Winnow/Perceptron.
- Others upcoming.

# Recommendation mining.

- Recommend items to users.



- Examples:
  - Find movies I might want to watch.
  - Find books related to the book I am buying.
  - Find people I might want to meet.
  - Recommend locations to items.

# Recommendation mining.

- Integrated Taste.
- Mature Java library.
- Java-based, web service / HTTP bindings.

- Batch mode based on EC2 and Hadoop.

# Frequent pattern mining

- Given groups of items, find commonly co-occurring items.

- Examples:
  - In shopping carts find items bought together.
  - In query logs find queries issued in one session.

**Release: 0.1**

Big Thanks to those who made this possible

Mahout is running on Amazon EMR.

**October 22, 2008 by e_calamar**
**http://www.flickr.com/photos/e_calamar/2964991182/**

# Why go for Apache Mahout?

# Jumpstart your project with proven code.

January 8, 2008 by dreizehn28
http://www.flickr.com/photos/1328/2176949559

# Discuss ideas and problems online.

**November 16, 2005 [phil h]**
**http://www.flickr.com/photos/hi-phi/64055296**

Become part of the community.

<project>-user@[lucene|hadoop].apache.org

<project>-dev@[lucene|hadoop].apache.org

July 9, 2006 by trackrecord
http://www.flickr.com/photos/trackrecord/185514449

Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.

# Sept., 29th 2009: Hadoop* Get Together in Berlin

- Thilo Götz: “JAQL”
- Thorsten Schütt: “Solving puzzles with Map/Reduce”
- Uwe Schindler: “Lucene 2.9 with focus on range search.”
- nugg.ad GmbH: “Using Hadoop for ad recommendation.”

newthinking store
Tucholskystr. 48

# December 2009: Hadoop* Get Together in Berlin.

* UIMA, Hbase, Lucene, Solr, katta, Mahout, CouchDB, pig, Hive, Cassandra, Cascading, JAQL, ... talks welcome as well.

<project>-user@[lucene|hadoop].apache.org

<project>-dev@[lucene|hadoop].apache.org

July 9, 2006 by trackrecord
http://www.flickr.com/photos/trackrecord/185514449

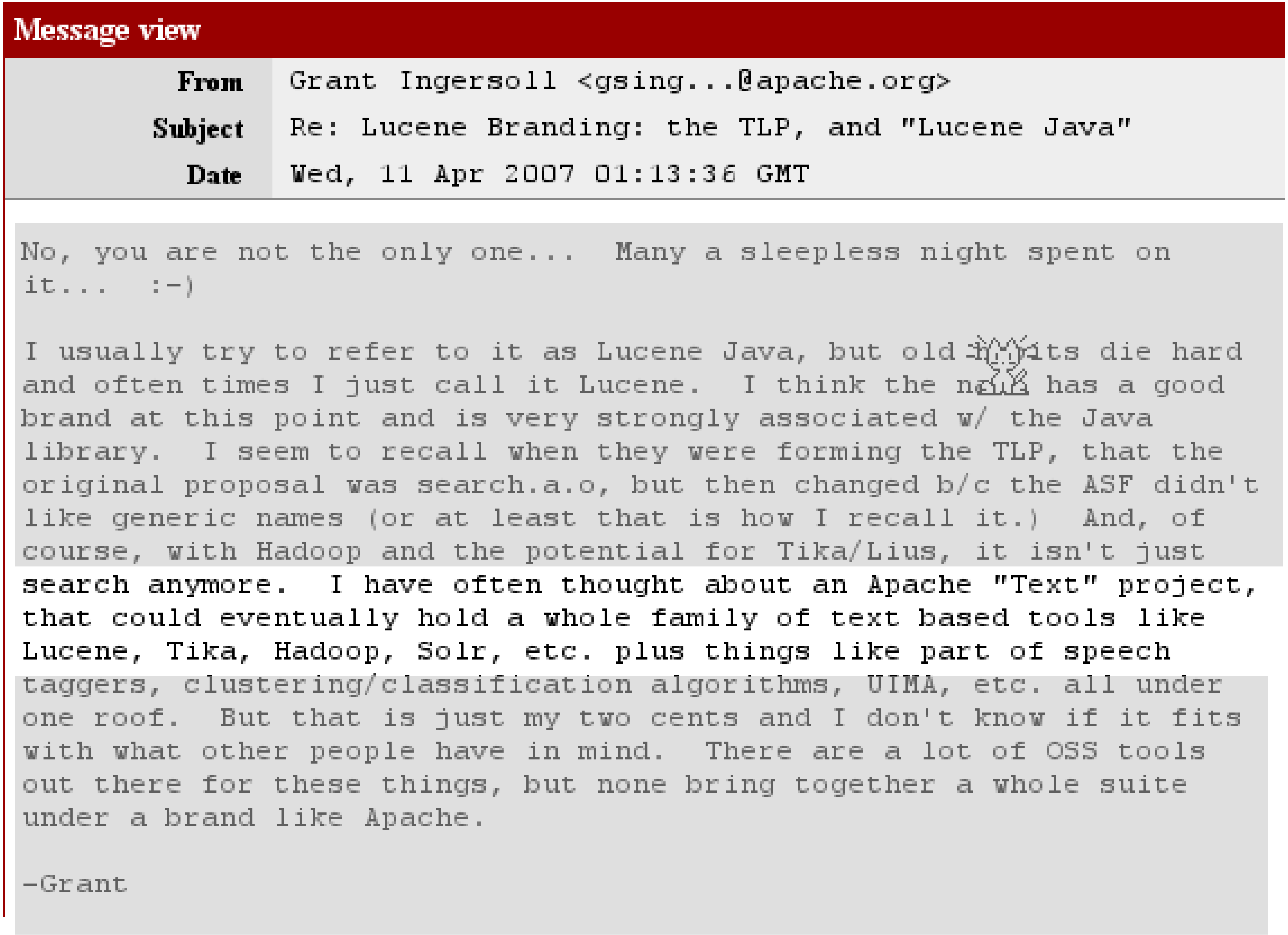
Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.



## Message view

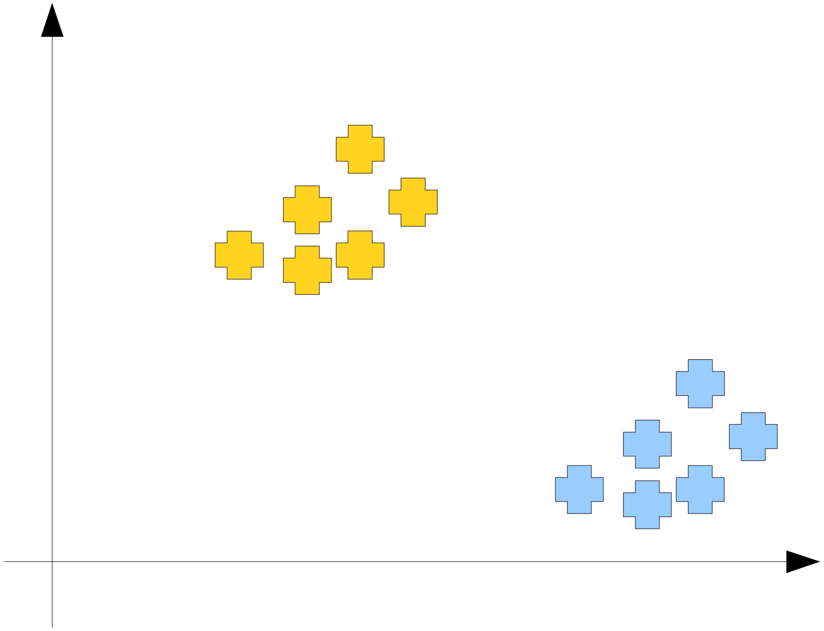
| | |
|---|---|
| **From** | Grant Ingersoll <gsing...@apache.org> |
| **Subject** | Re: Lucene Branding: the TLP, and "Lucene Java" |
| **Date** | Wed, 11 Apr 2007 01:13:36 GMT |

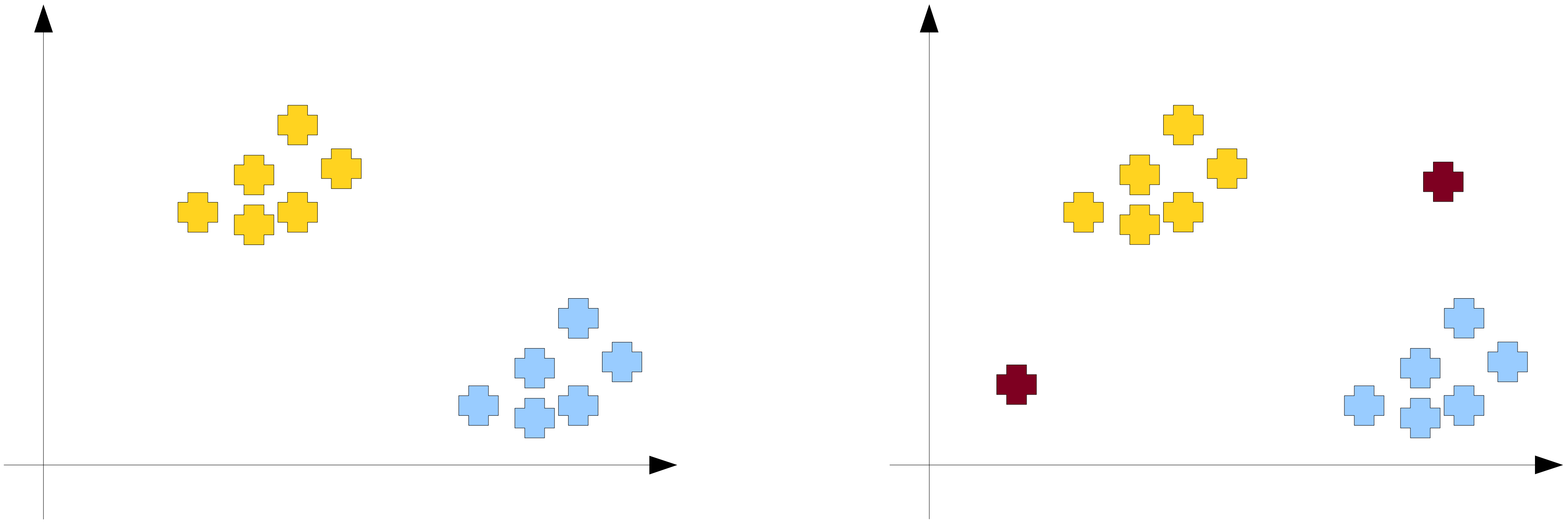
No, you are not the only one... Many a sleepless night spent on it... :-)

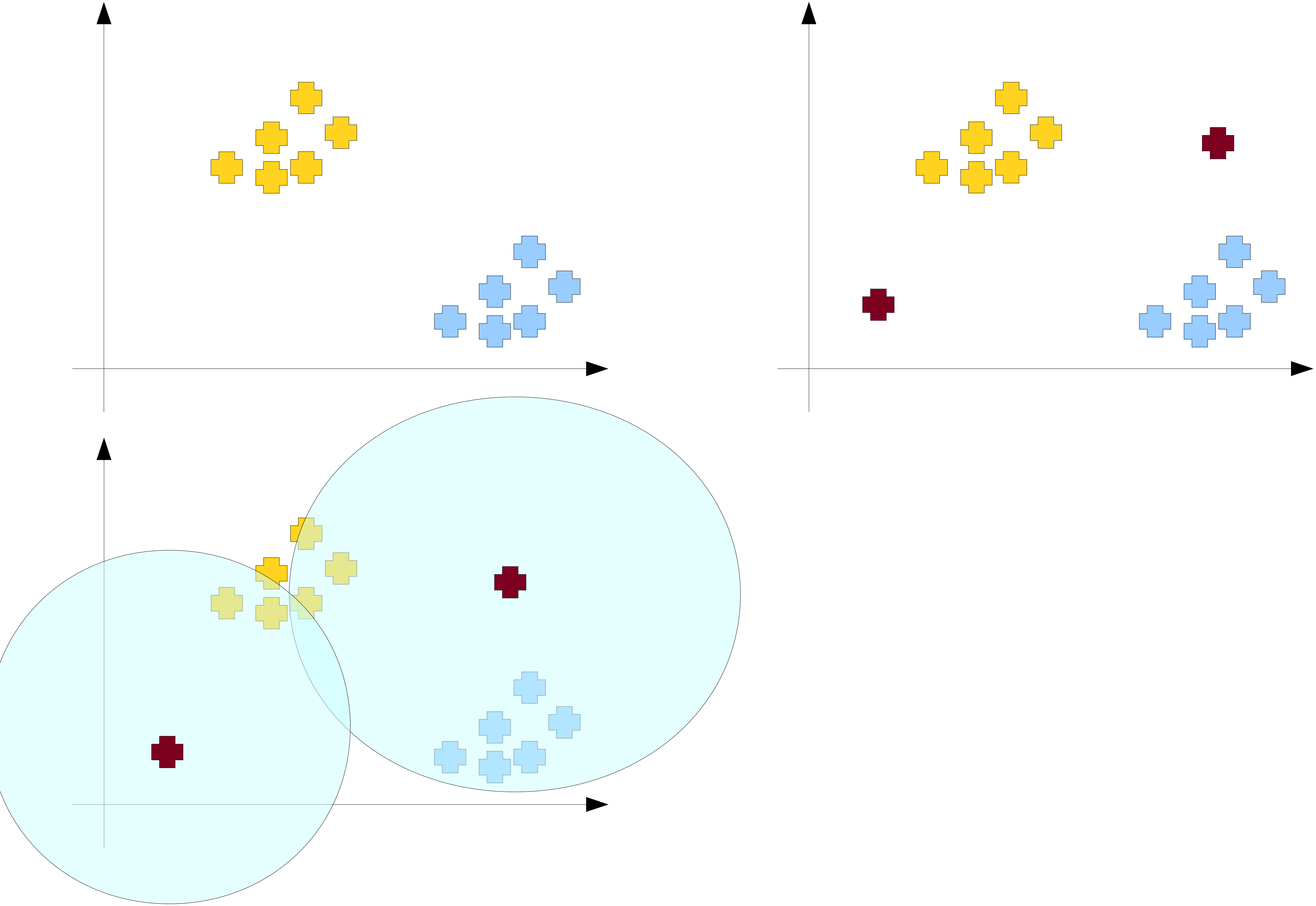
I usually try to refer to it as Lucene Java, but old habits die hard and often times I just call it Lucene. I think the name has a good brand at this point and is very strongly associated w/ the Java library. I seem to recall when they were forming the TLP, that the original proposal was search.a.o, but then changed b/c the ASF didn't like generic names (or at least that is how I recall it.) And, of course, with Hadoop and the potential for Tika/Lius, it isn't just search anymore. I have often thought about an Apache "Text" project, that could eventually hold a whole family of text based tools like Lucene, Tika, Hadoop, Solr, etc. plus things like part of speech taggers, clustering/classification algorithms, UIMA, etc. all under one roof. But that is just my two cents and I don't know if it fits with what other people have in mind. There are a lot of OSS tools out there for these things, but none bring together a whole suite under a brand like Apache.

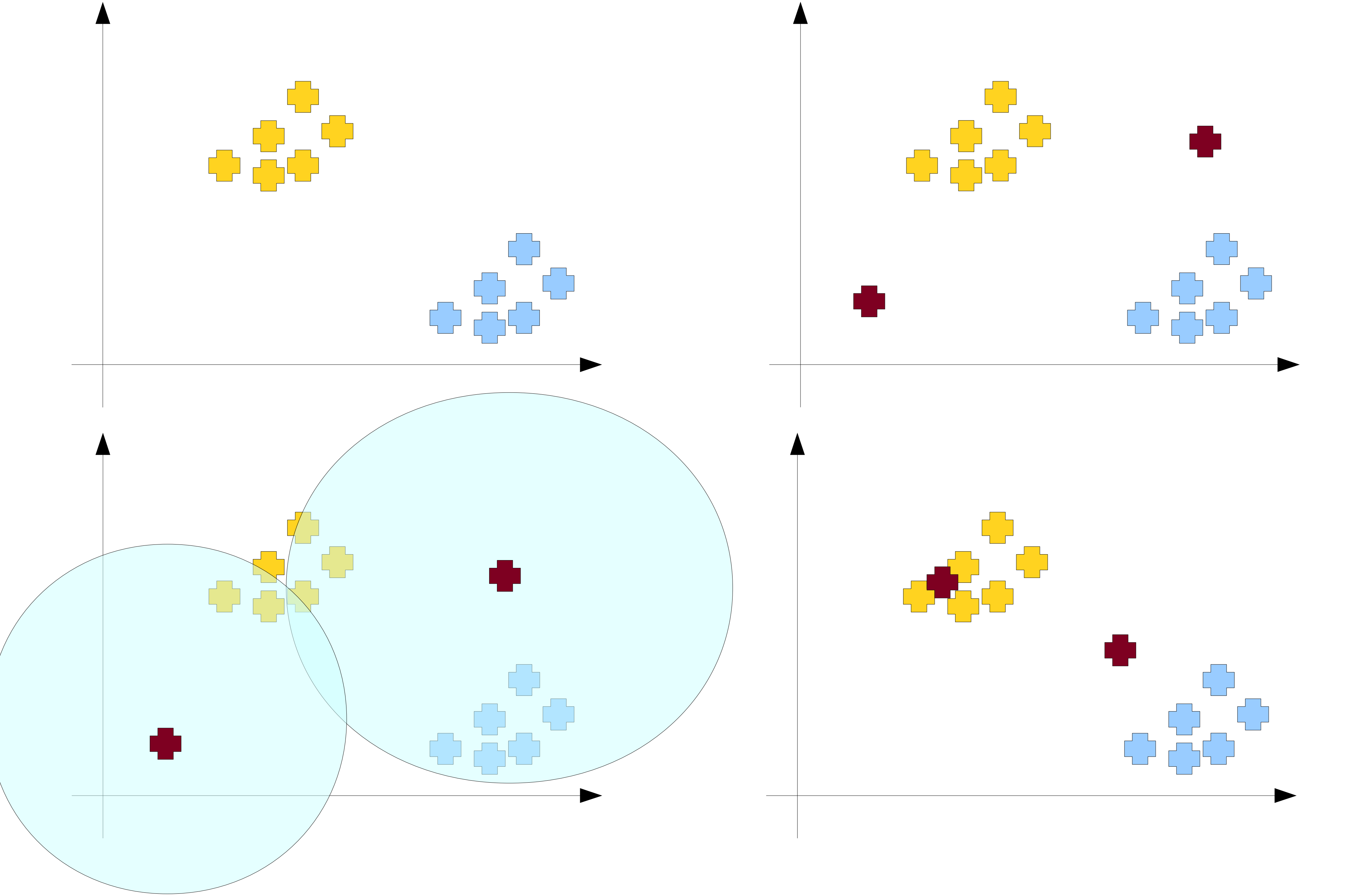
-Grant

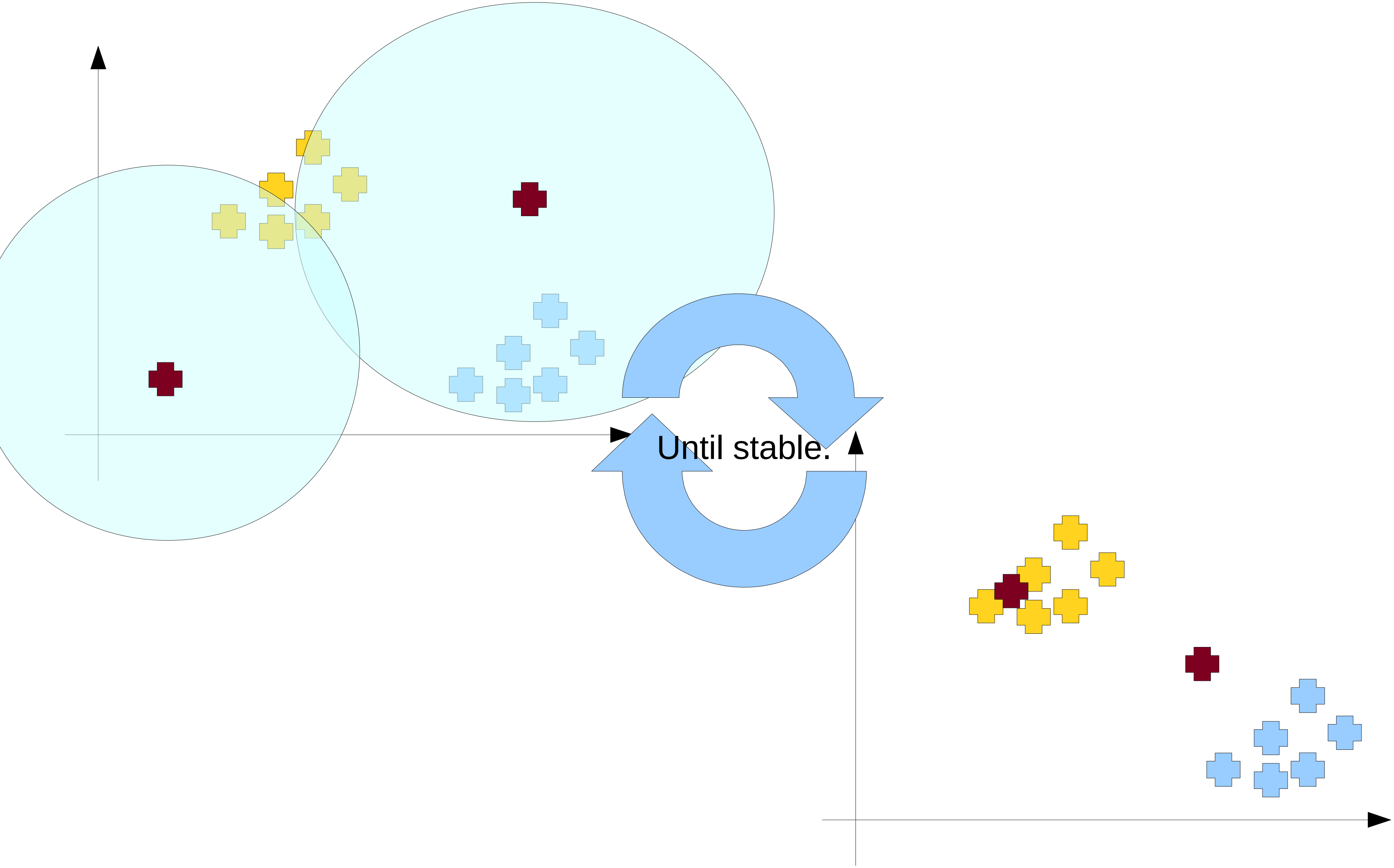
# Going parallel: k-Means



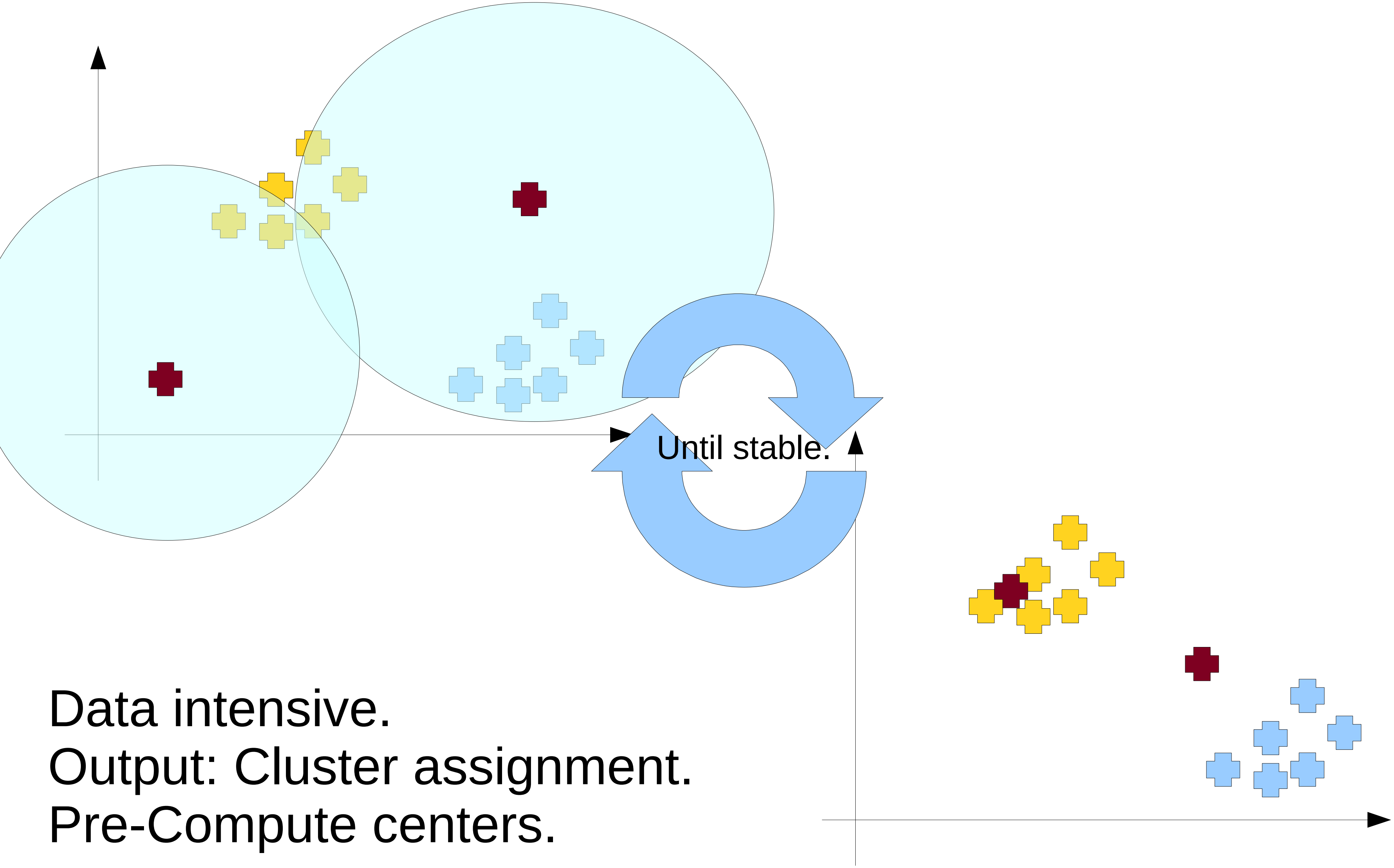






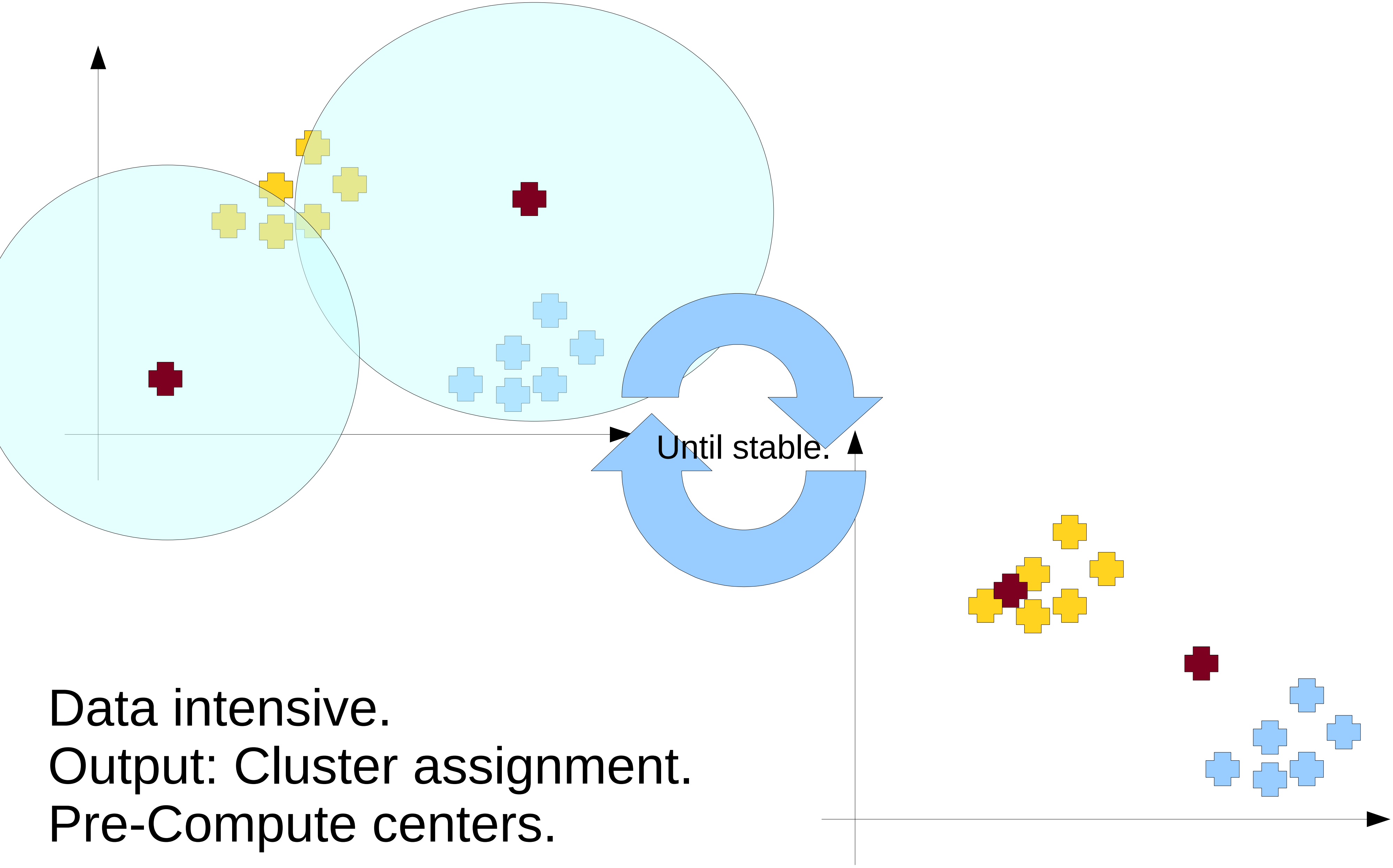


Until stable.



Data intensive.
Output: Cluster assignment.
Pre-Compute centers.

Done in Map.



Data intensive.
Output: Cluster assignment.
Pre-Compute centers.

Done in Map.

Done in Reduce.