Apache Mahout

Making data analysis easy

Isabel Drost

Nighttime:

Co-Founder, committer Apache Mahout.
Organiser of Berlin Hadoop Get Together.

Daytime:

Software developer.
Guest lecturer at TU Berlin.
Co-Organiser Berlin Buzzwords 2010.

Hello codebits!

Hello codebits!

Hello codebits!

Hello codebits!

Hello codebits!

Machine learning
background?

Hello codebits!

Hello codebits!

Agenda

- Data Mining/ Machine Learning?
- Why is scaling hard?
- Introducing Apache Hadoop.
- Going beyond simple statistics.

Machine learning – what's that?

Data Mining Applications

- Marketing.
- Surveillance.
- Fraud Detection.
- Scientific Discovery.
- Discover items usually purchased together.

= Extracting patterns from data.

Machine Learning Applications

- E-Mail spam classification.
- News-topic discovery.
- Building recommender systems.

= Extracting prediction models from data.

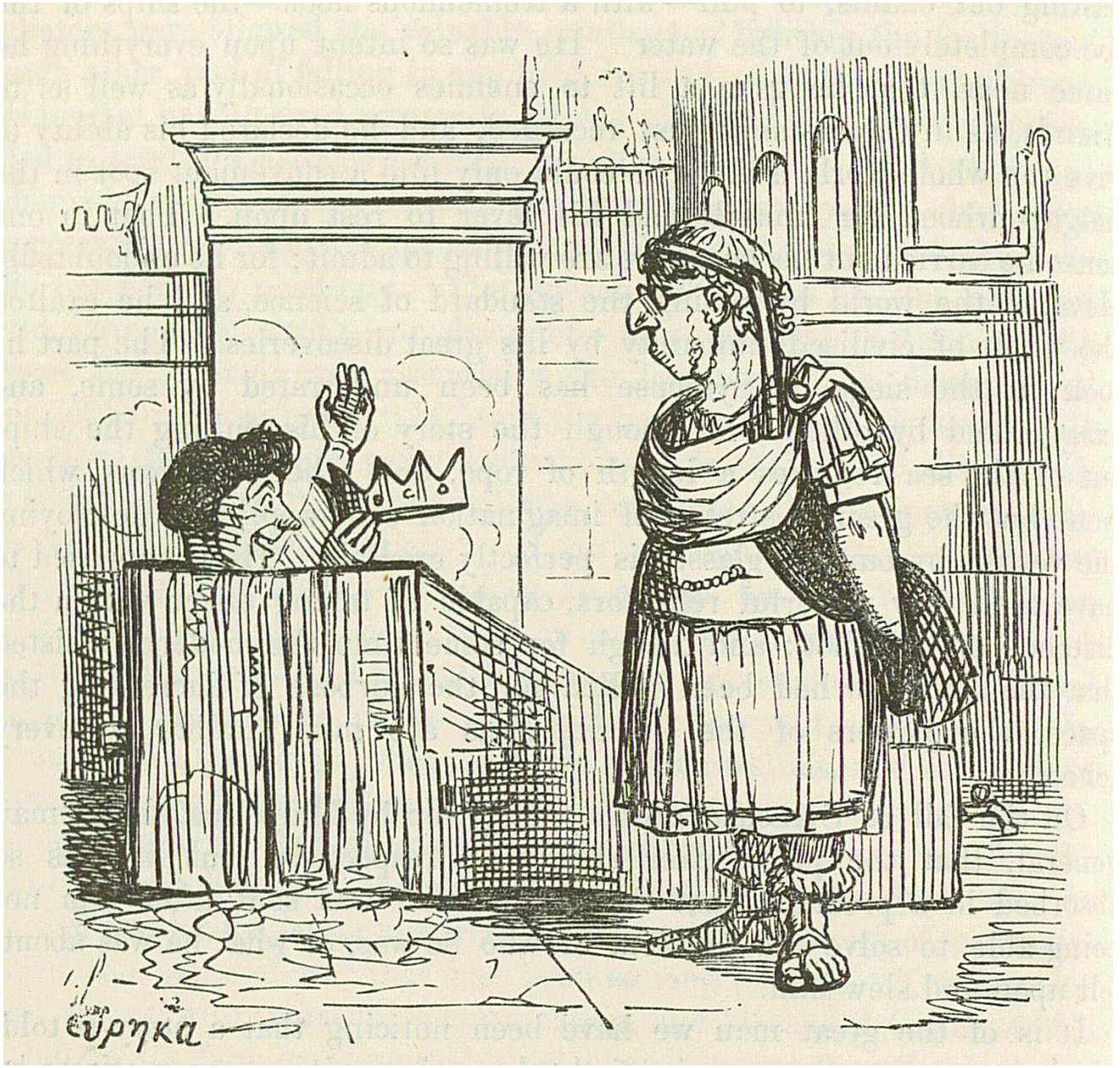


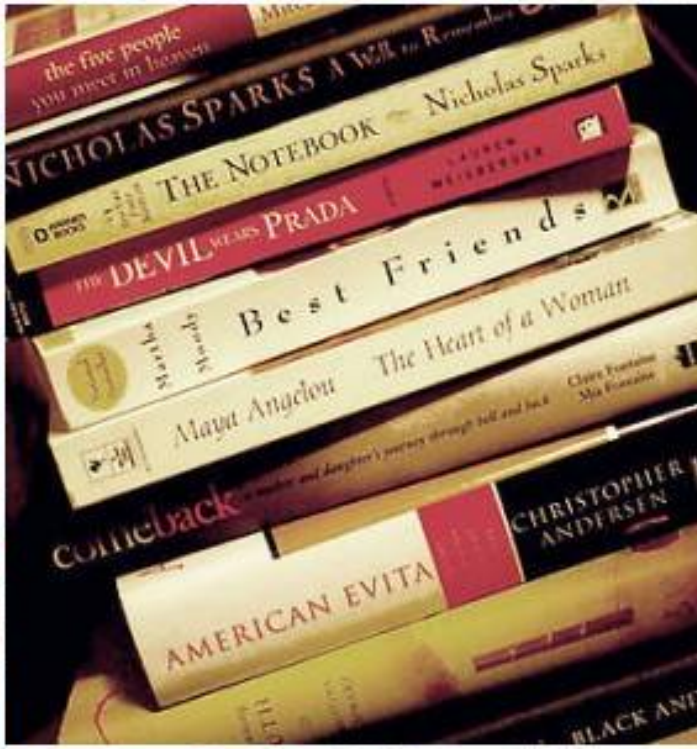
Image by John Leech, from: The Comic History of Rome by Gilbert Abbott A Beckett.
Bradbury, Evans & Co, London, 1850s
Archimedes taking a Warm Bath

Archimedes model of nature

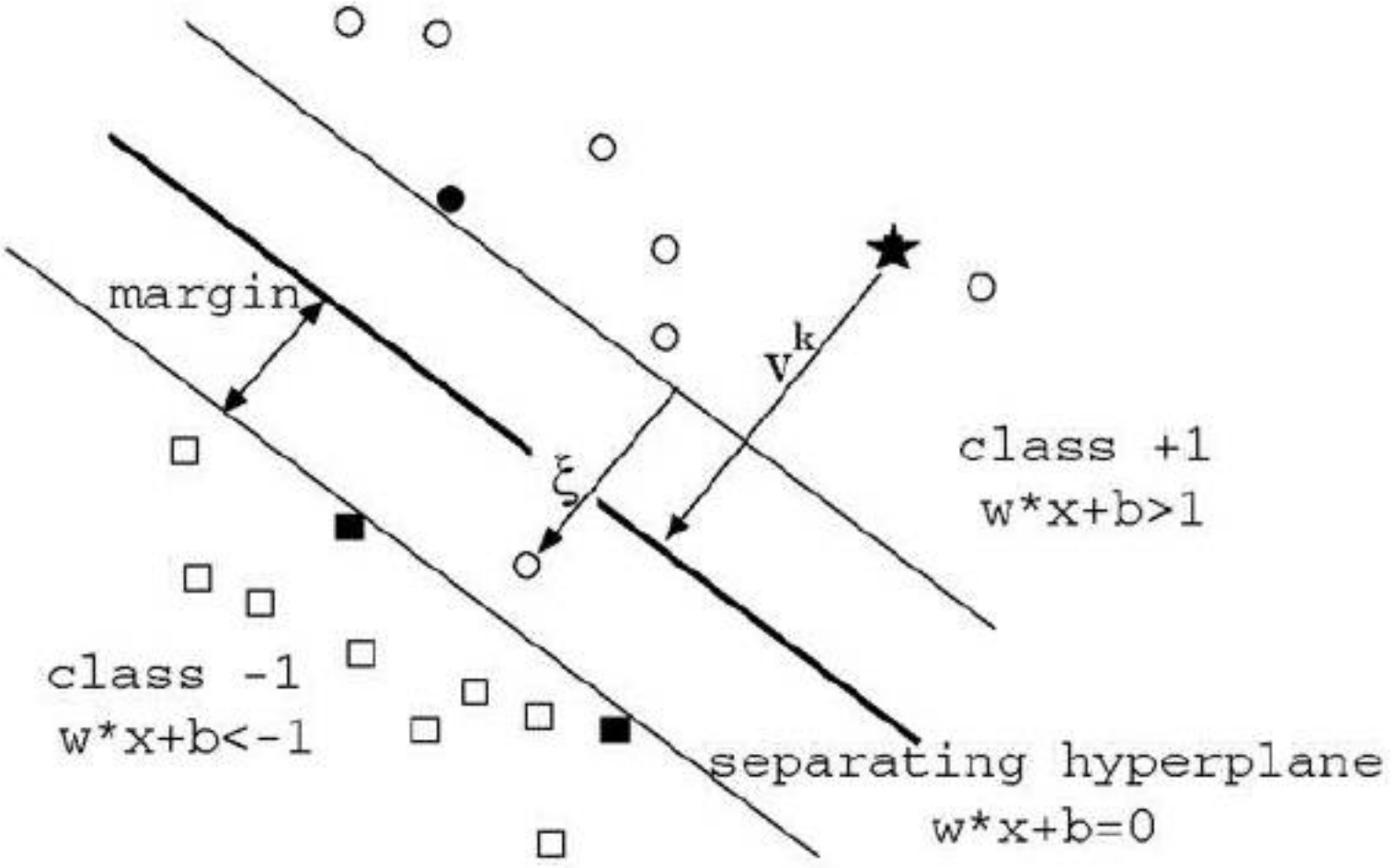
$$\frac{Density\ of\ Object}{Density\ of\ Fluid} = .$$

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

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



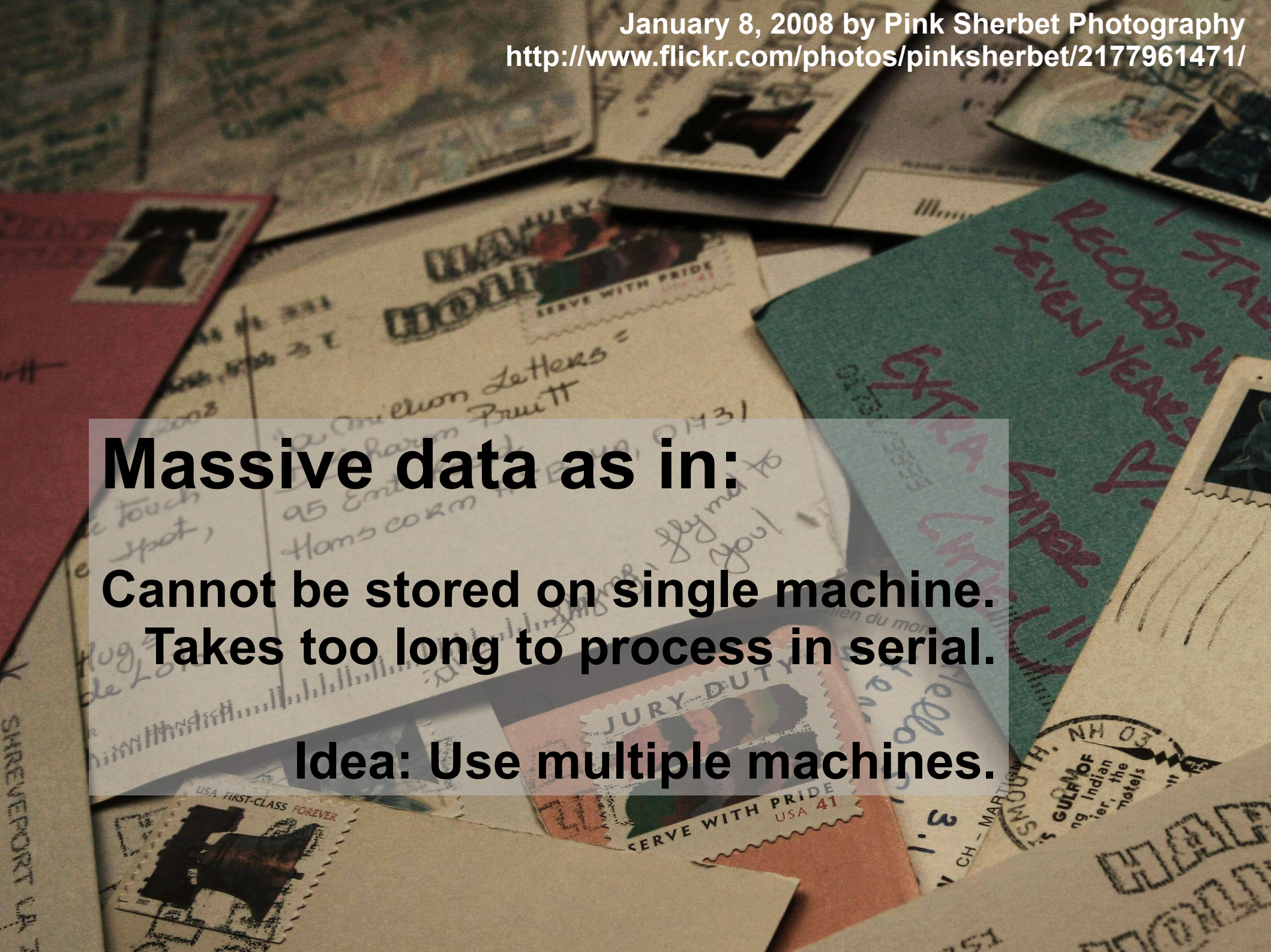
An SVM's model of nature



The challenge

Mission

Provide scalable data mining algorithms.


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.

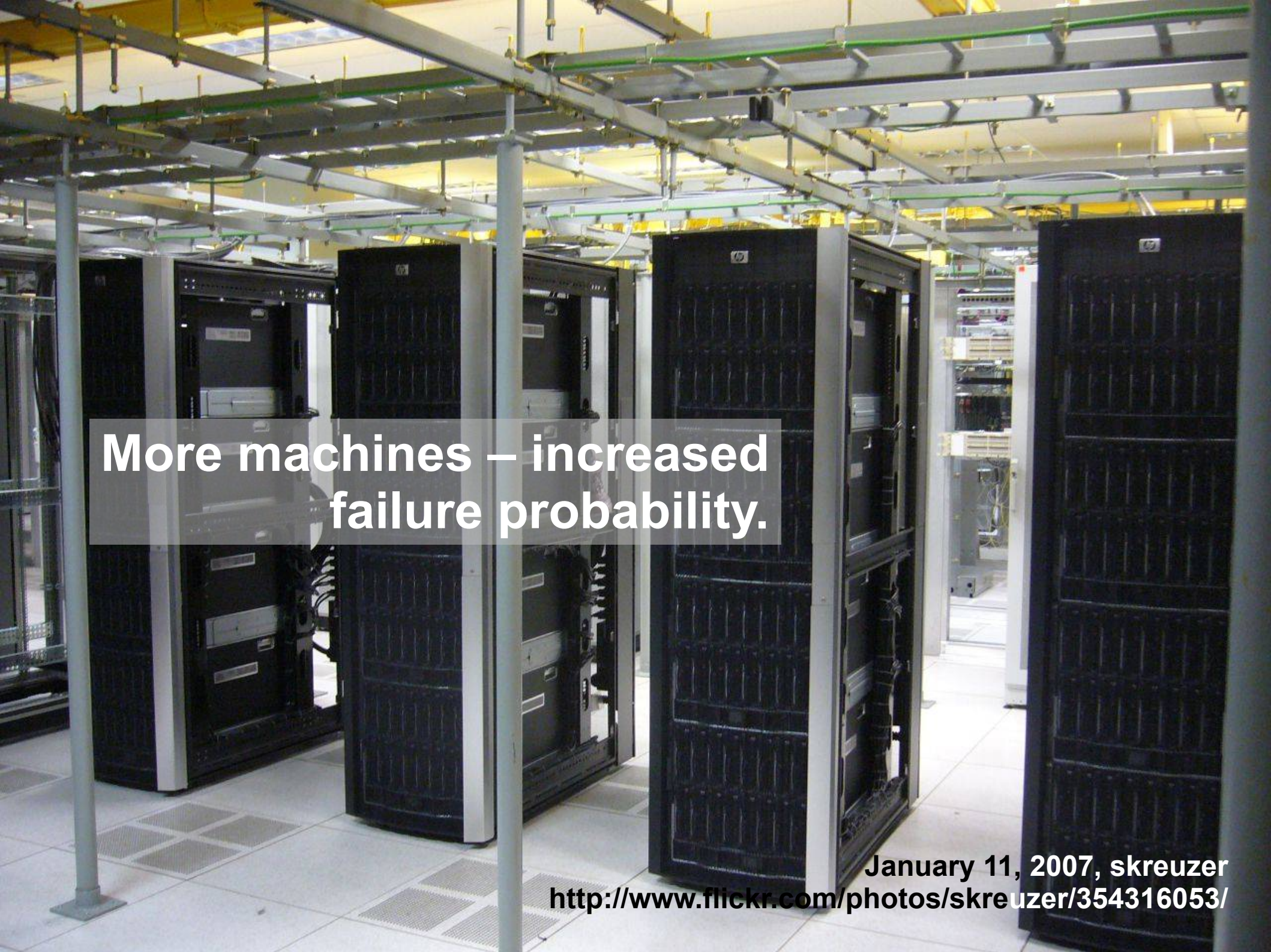
Challenges when scaling out.

Single machines tend to fail:

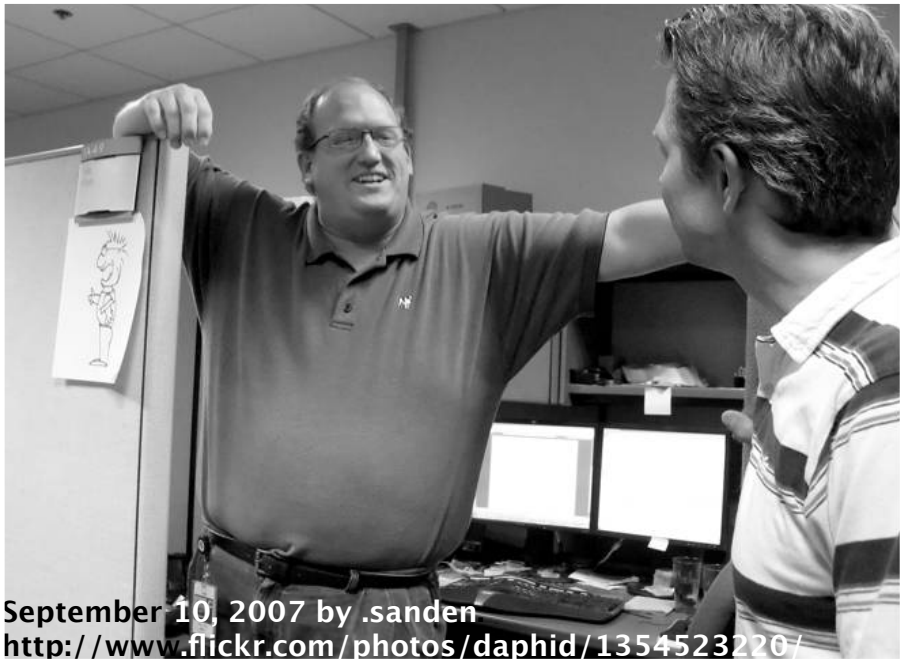
Hard disk.

Power supply.

...


More machines – increased failure probability.
January 11, 2007, skreuzer
http://www.flickr.com/photos/skreuzer/354316053/

Typical developer



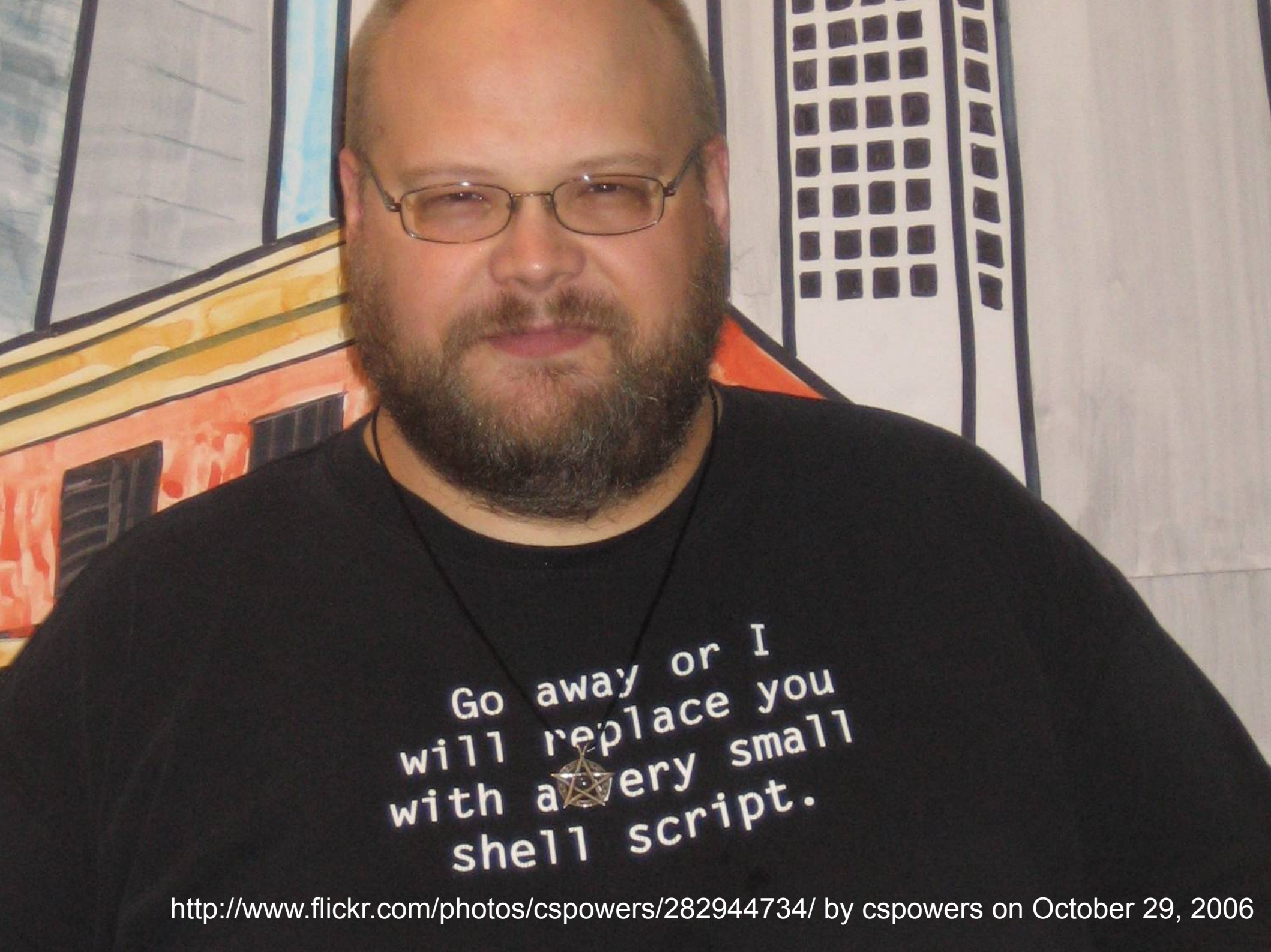
- Has never dealt with large (petabytes) amount of data.
- Has no thorough understanding of parallel programming.
- Has no time to make software production ready.

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

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

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



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

Easy distributed programming.

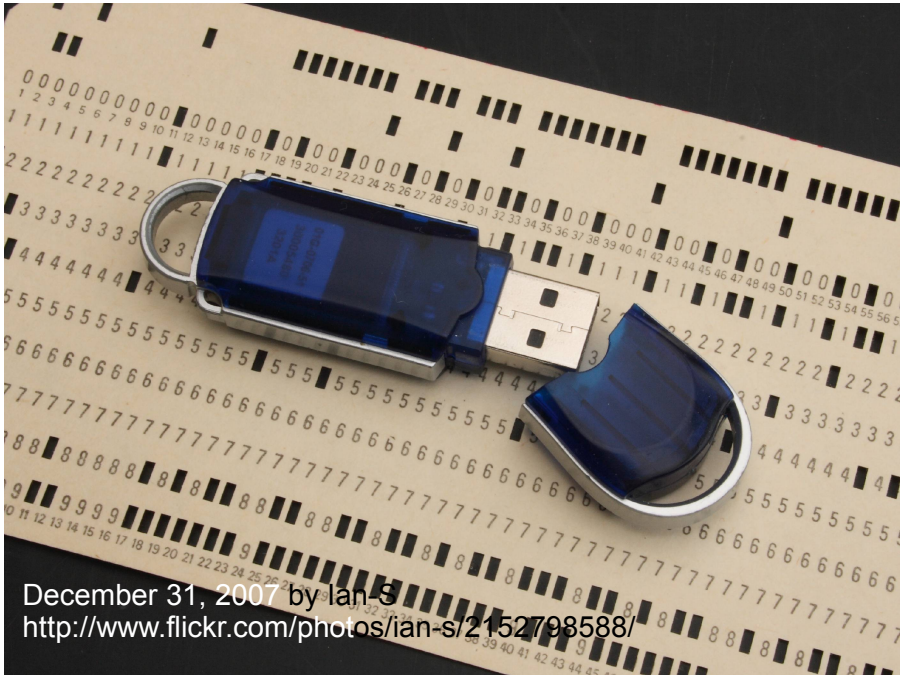
Well known in industry and research.

Scales well beyond 1000 nodes.

Hadoop assumptions

Assumptions:

Moving computation is cheap.
Moving data is expensive.



Ideas:

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

Assumptions:

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

August 27, 2007 by Stuart Bryant
http://www.flickr.com/photos/stuartbryant/1249649001/

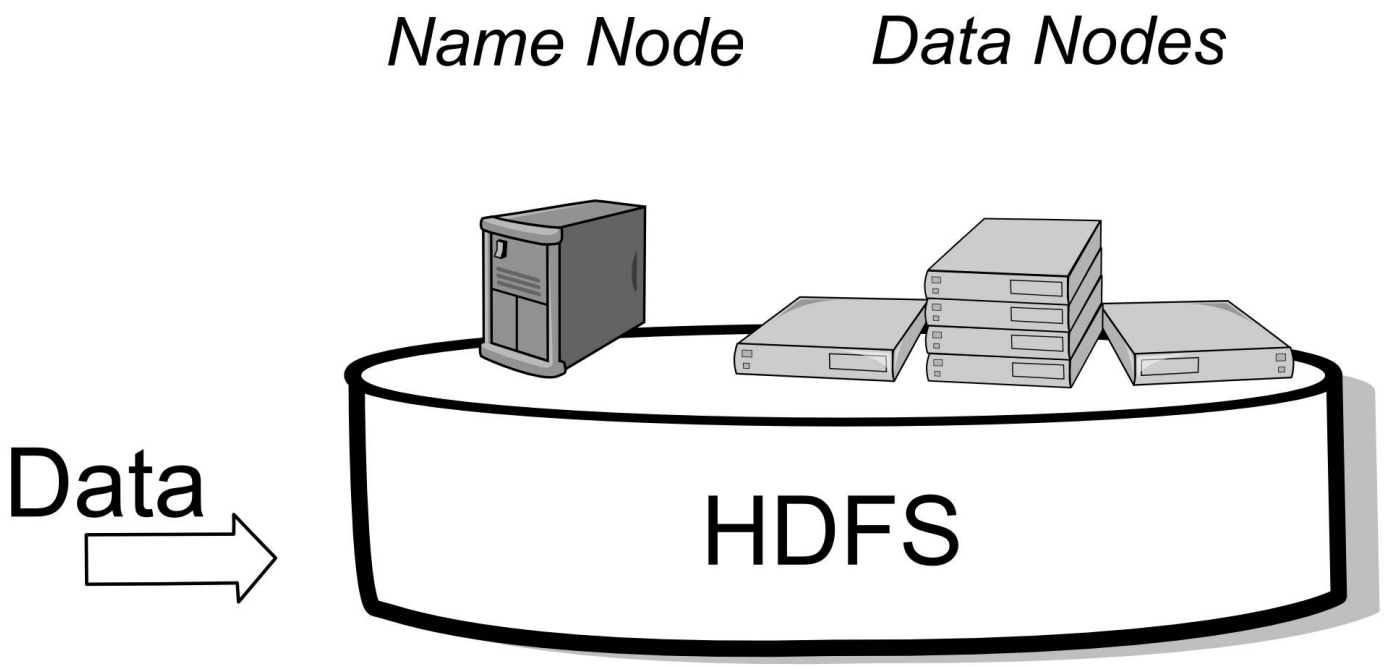
Ideas:

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


hadoop
HDFS


hadoop
MapReduce

HDFS building blocks

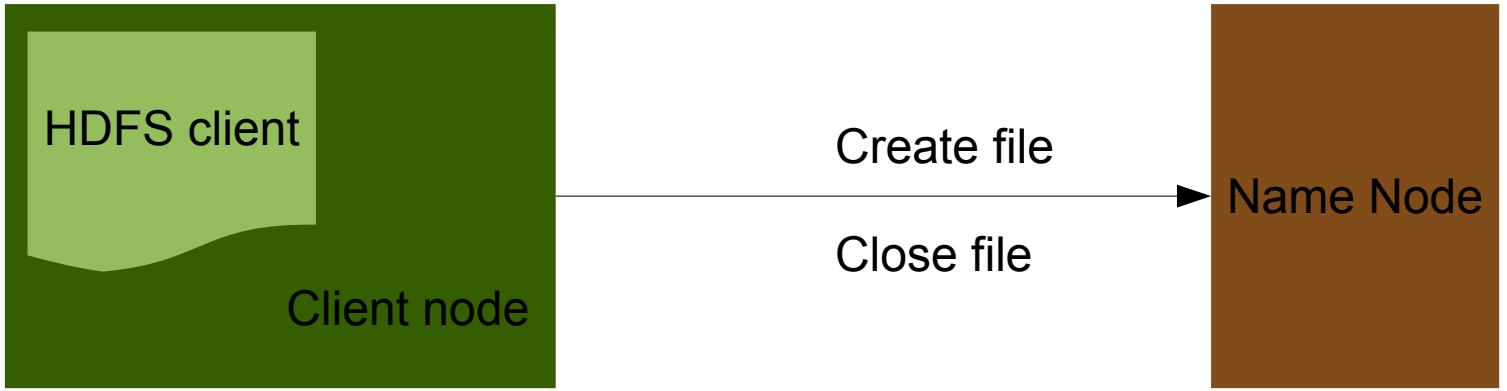


(Graphics: Thanks to Thilo.)

Anatomy of a file write

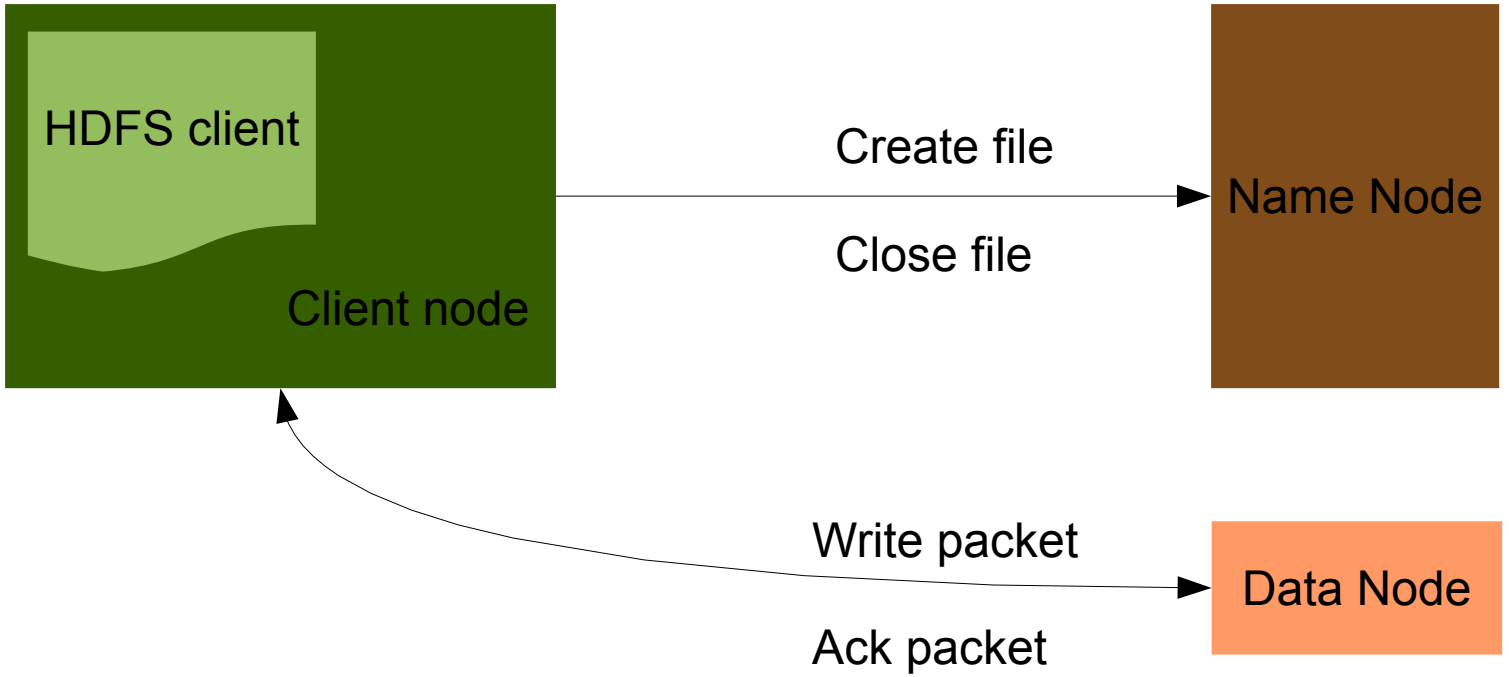
Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a file write



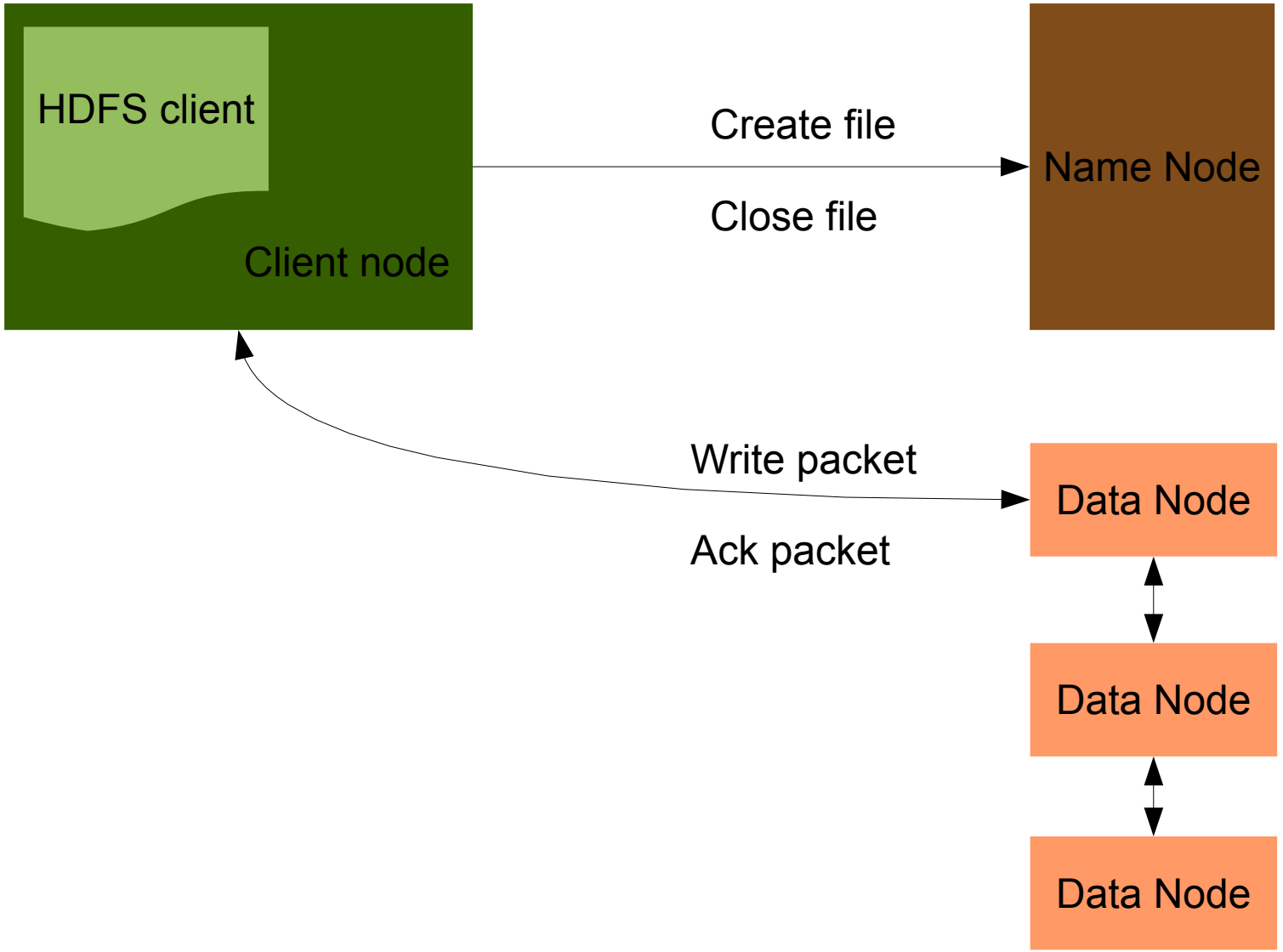
Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a file write



Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a file write



Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

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 Intelli
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
```

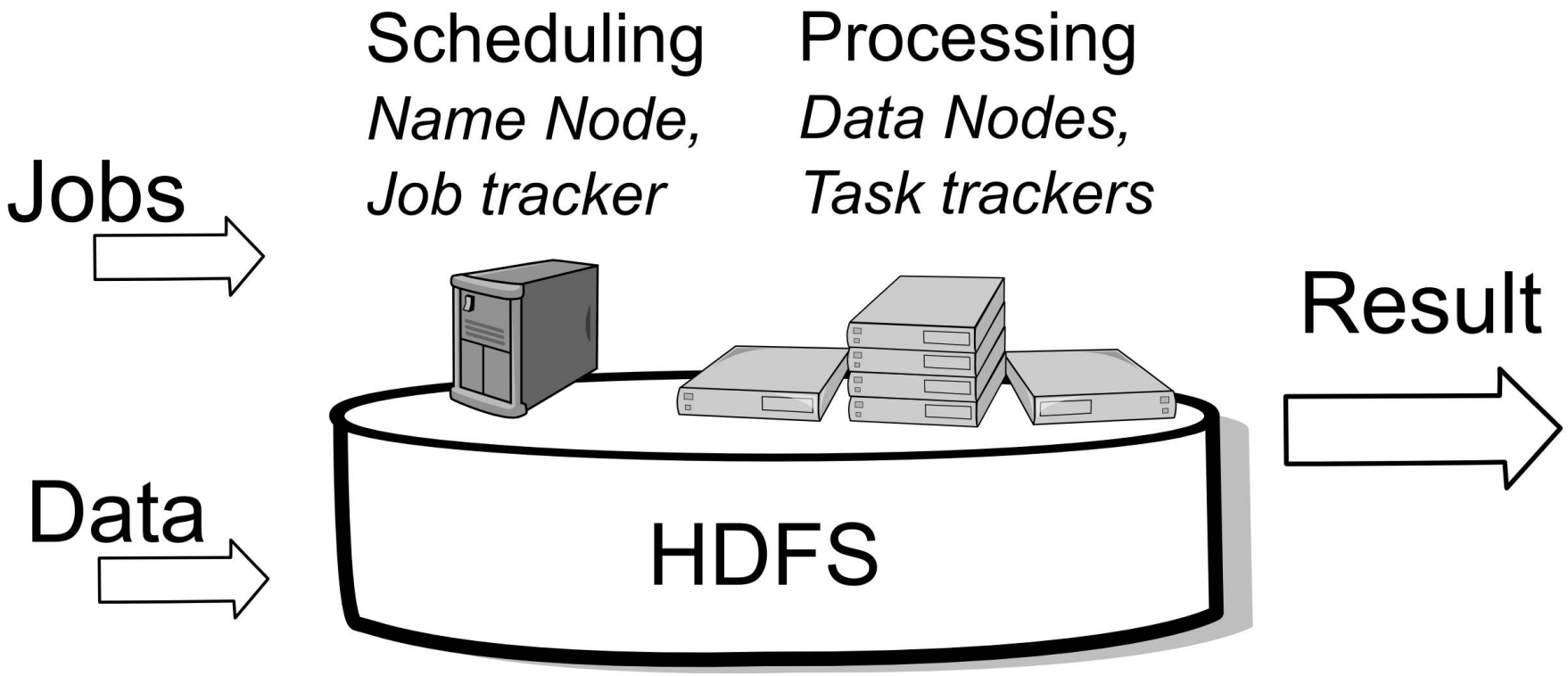
```
pattern=”http://[0-9A-Za-z\-_\.]*”

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

pattern=”http://[0-9A-Za-z\-_\.]*”

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

 M  A   P
```

```
| sort


| SHUFFLE
```

```
| uniq --count


| R E D U C E
```



(Graphics: Thanks to Thilo.)

Mission

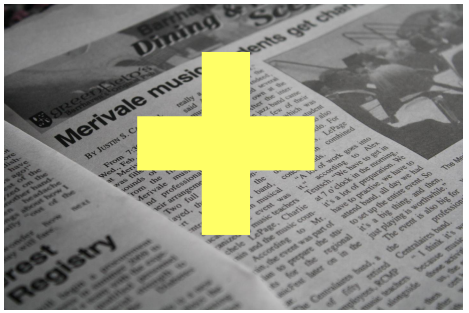
Provide scalable data mining algorithms.

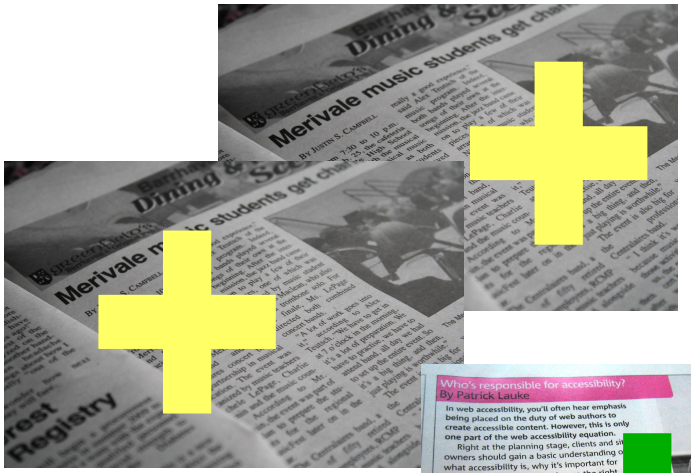
HowTo: From data to information.

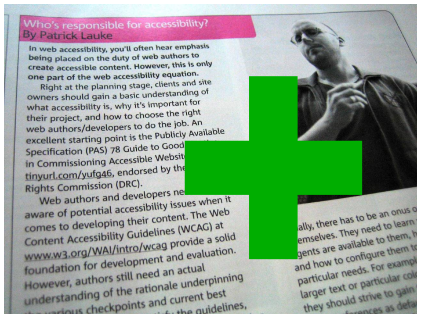
January 3, 2006 by Matt Callow
http://www.flickr.com/photos/blackcustard/81680010

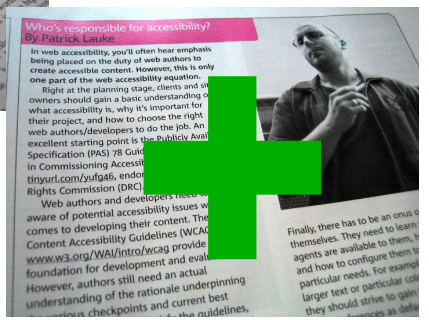
http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

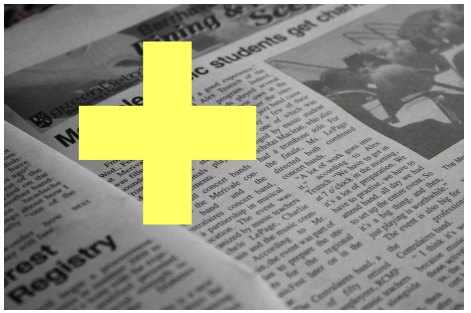
http://www.flickr.com/photos/redux/409356158/

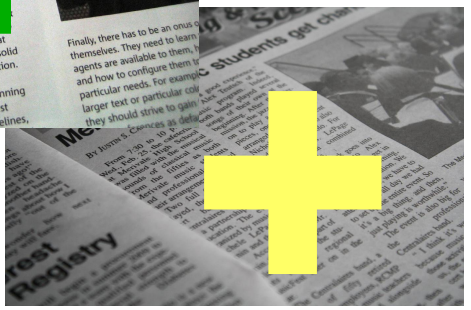












http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

http://www.flickr.com/photos/redux/409356158/in/photostream/

http://www.flickr.com/photos/topsy/204929063/

ZEIT ONLINE | DATENSCHUTZ

Partnersuche

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIU
SPORT

Internet | Datenschutz | Mobil | Games

DATENSCHUTZ

Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.

© Tim Brakemeier/dpa

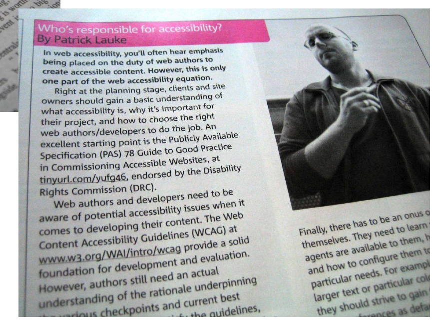
In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz

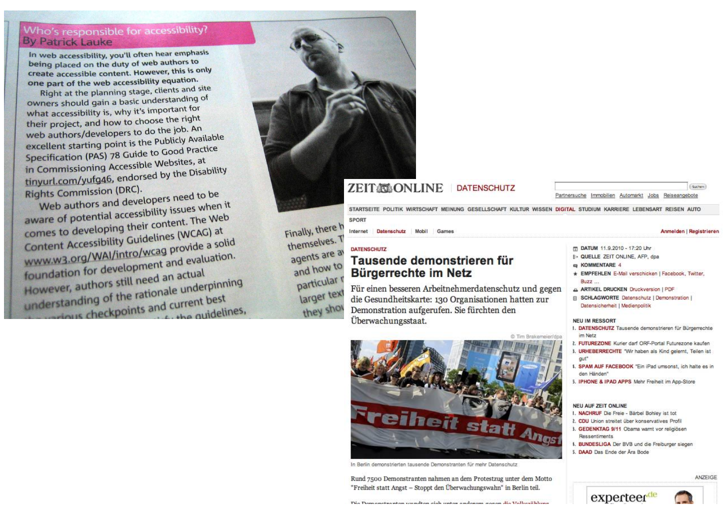
Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

ANZEIGE

http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

http://www.flickr.com/photos/redux/409356158/





From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.

DATENSCHUTZ

Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.

© Tim Brakemeier/dpa

In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz

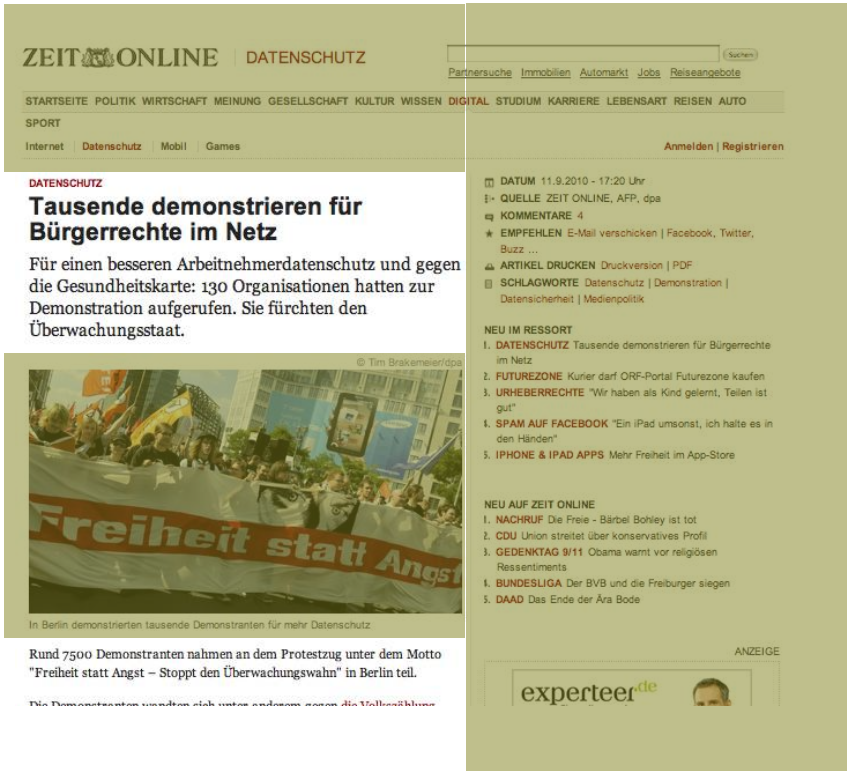
Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

Die Demonstranten wandten sich unter anderem gegen die Volkszählung

DATUM 11.9.2010 - 17:20 Uhr
QUELLE ZEIT ONLINE, AFP, dpa
KOMMENTARE 4
EMPFEHLEN E-Mail verschicken | Facebook, Twitter, Buzz ...
ARTIKEL DRUCKEN Druckversion | PDF
SCHLAGWORTE Datenschutz | Demonstration | Datensicherheit | Medienpolitik

NEU IM RESSORT
1. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte im Netz
2. FUTUREZONE Kurier darf ORF-Portal Futurezone kaufen
3. URHEBERRECHTE "Wir haben als Kind gelernt, Teilen ist gut"
4. SPAM AUF FACEBOOK "Ein iPad umsonst, ich halte es in den Händen"
5. IPHONE & IPAD APPS Mehr Freiheit im App-Store

NEU AUF ZEIT ONLINE
1. NACHRUF Die Freie - Bärbel Bohley ist tot
2. CDU Union streitet über konservatives Profil
3. GEDENKTAG 9/11 Obama warnt vor religiösen Ressentiments
4. BUNDESLIGA Der BVB und die Freiburger siegen
5. DAAD Das Ende der Ära Bode

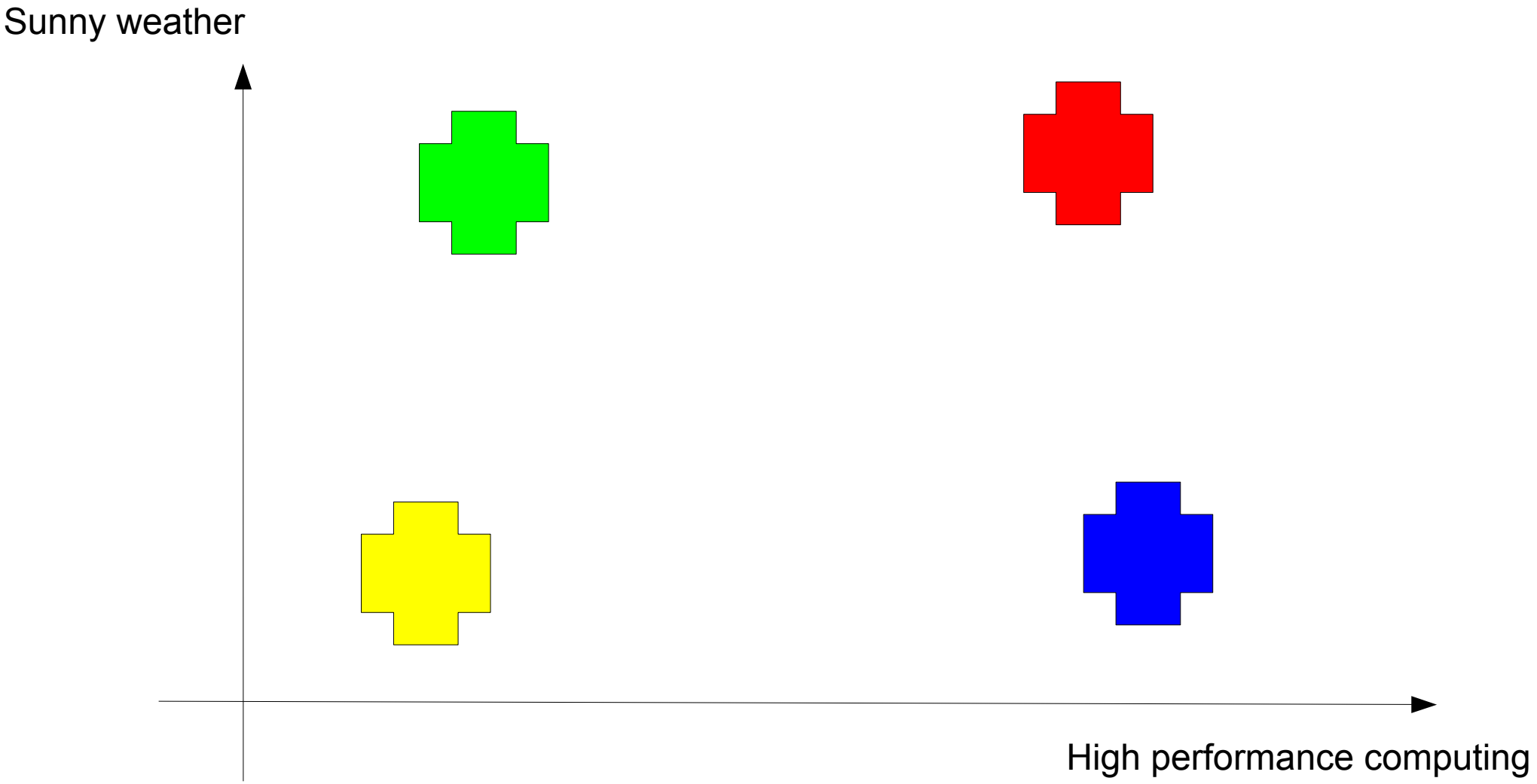


- Remove noise.

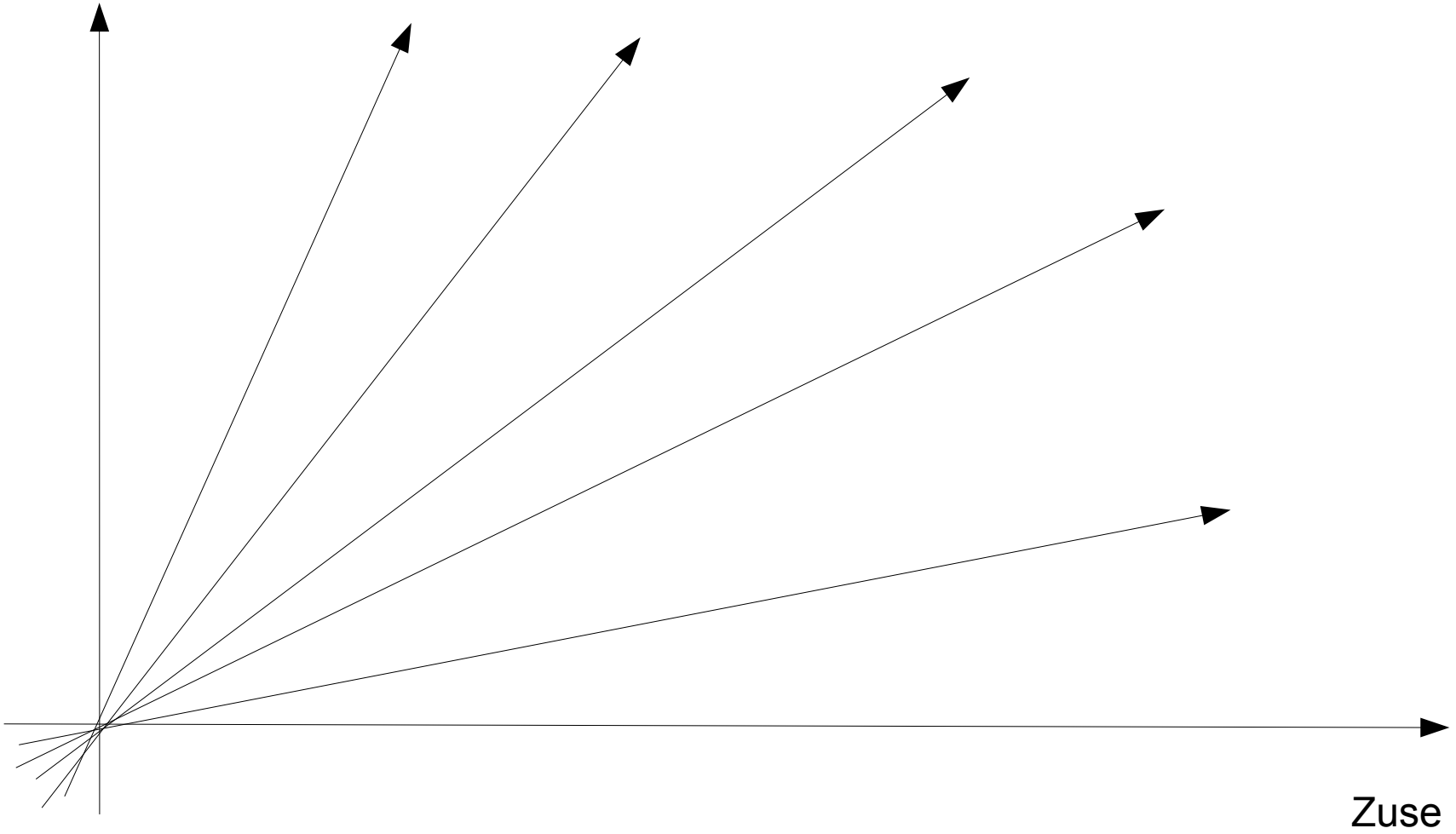
- Remove noise.
- Convert text to vectors.

From texts to vectors

If we looked at two words only:



Aaron



Binary bag of words

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Entry in vector is one, if word occurs in text.

$$b_{i,j} = \begin{cases} 1 \; \forall \; x_i \in d_j \\ 0 \; else \end{cases}$$

- Problem:
  - Number of word occurrences not accounted for.

Term Frequency

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Entry in vector equal to the words frequency.

$$b_{i,j} = n_{i,j}$$

- Problem:
  - Common words dominate vectors.

TF with stop wording

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the words frequency.

$$b_{i,j} = n_{i,j}$$

- Problem:
  - Common and uncommon words with same weight.

TF- IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.
- Problem:
  - Long texts get larger values.

$$b_{i,j} = n_{i,j} \times \log\left(\frac{|D|}{|\{d : t_i \in d\}|}\right)$$

Normalized TF- IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.
- Normalize vectors.
- Problem:
  - Additional domain knowledge ignored.

$$b_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log\left(\frac{|D|}{|\{d : t_i \in d\}|}\right)$$

Reality

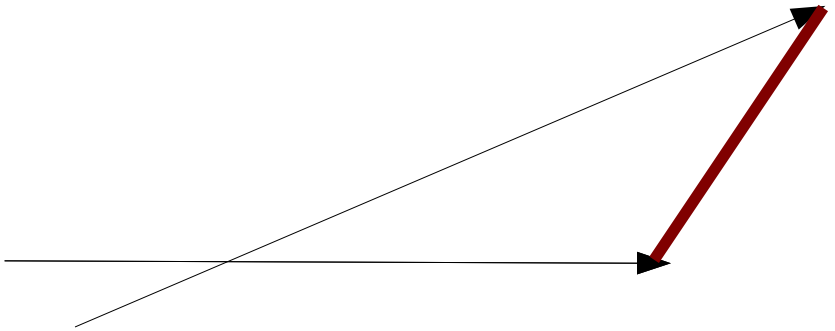
- There are a few more words in news.
- Use all relevant features/ signals available.
  - Words.
  - Header fields.
  - Characteristics of publishing url.
  - …
- Usually pipeline of feature extractors.

From data to information.

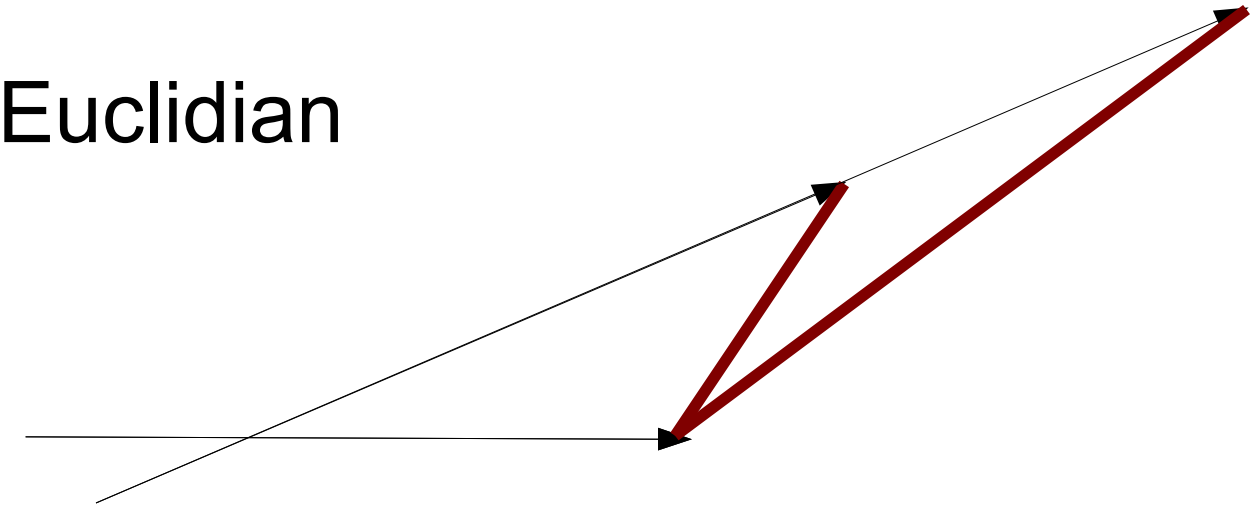
- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.

Step 2: Similarity

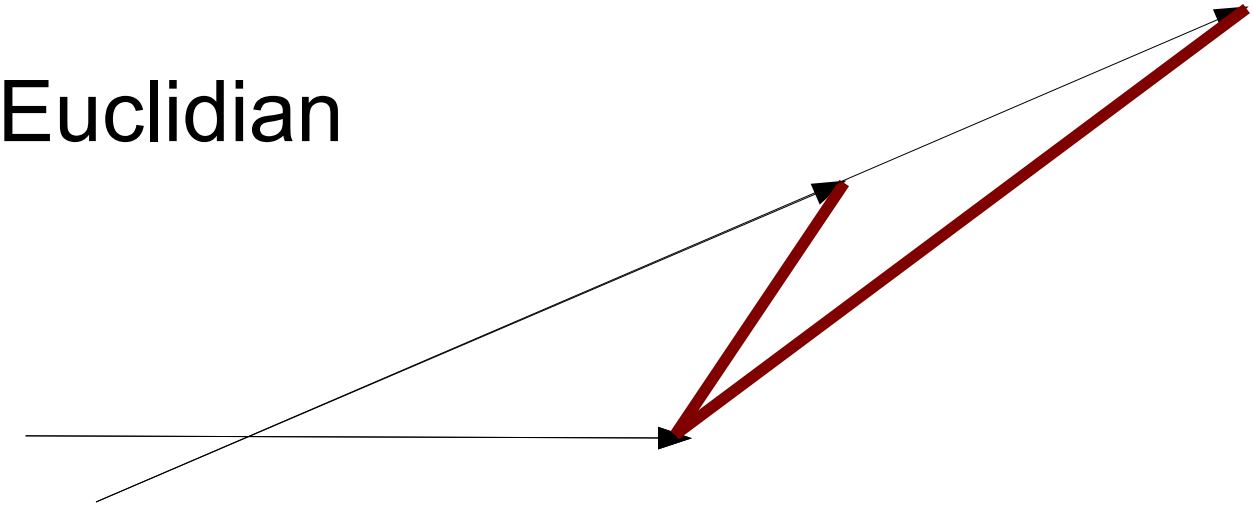
Euclidian



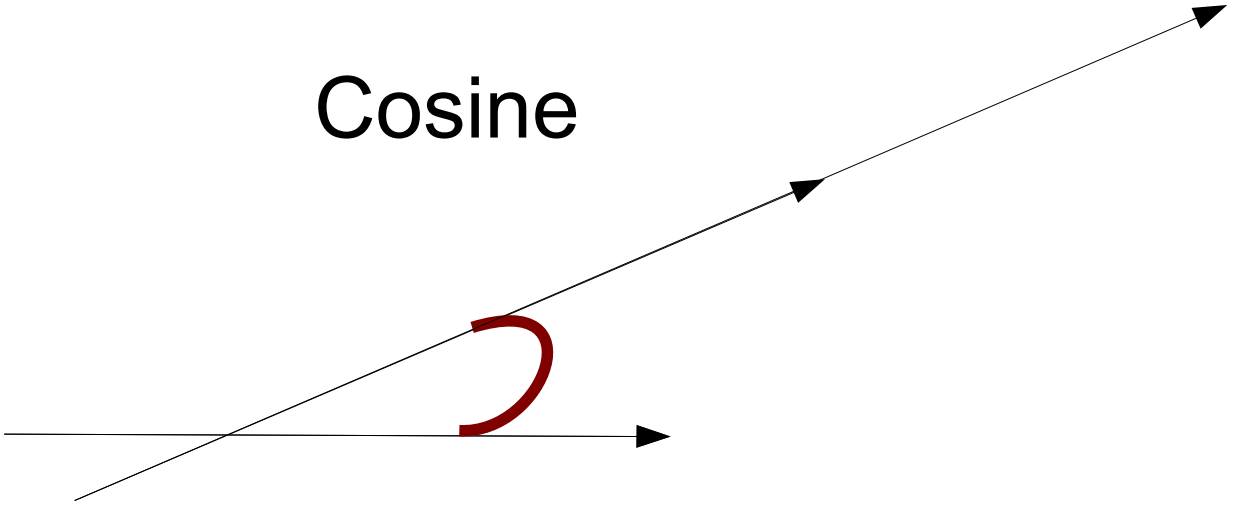
Euclidian



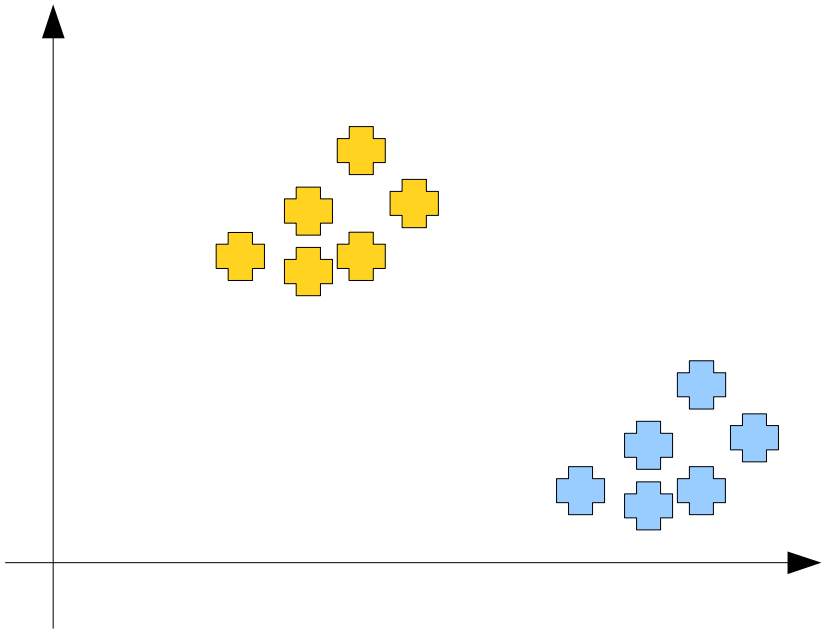
Euclidian

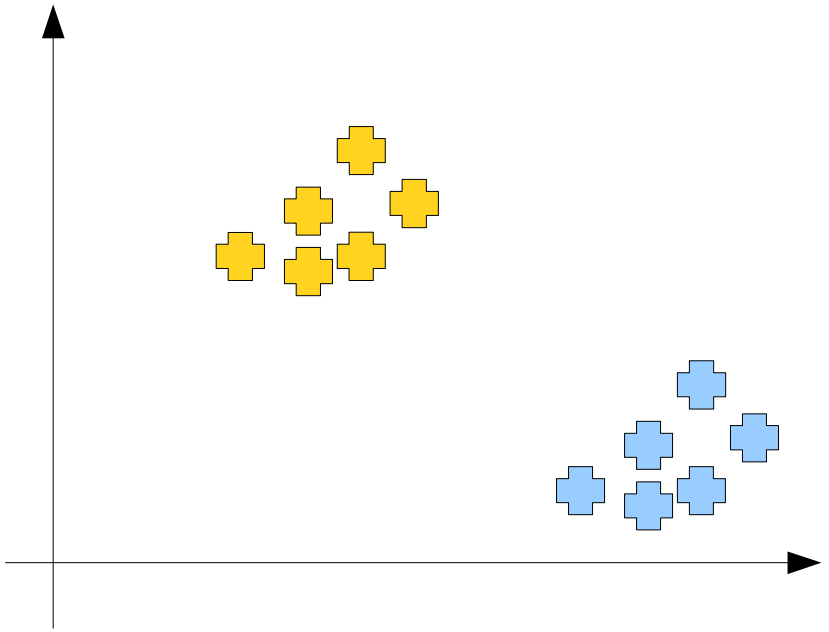


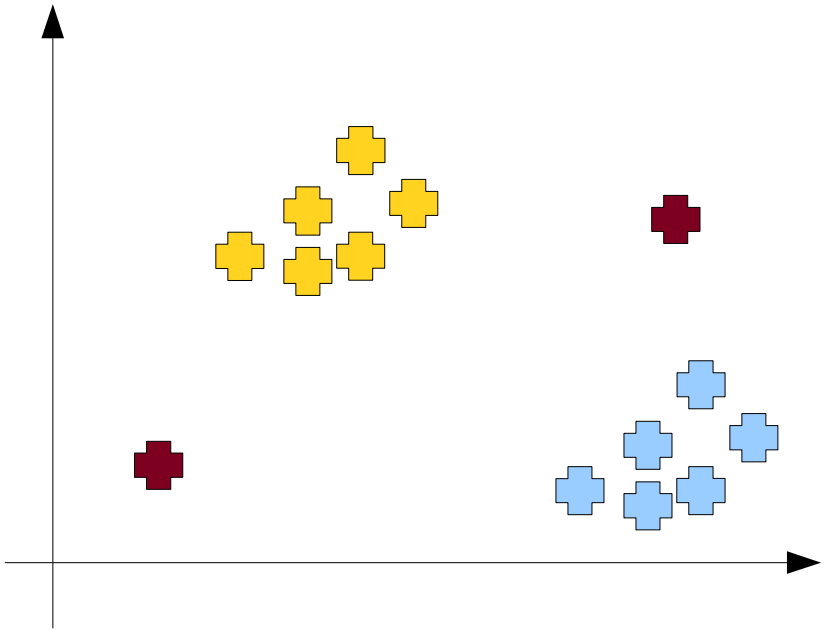
Cosine

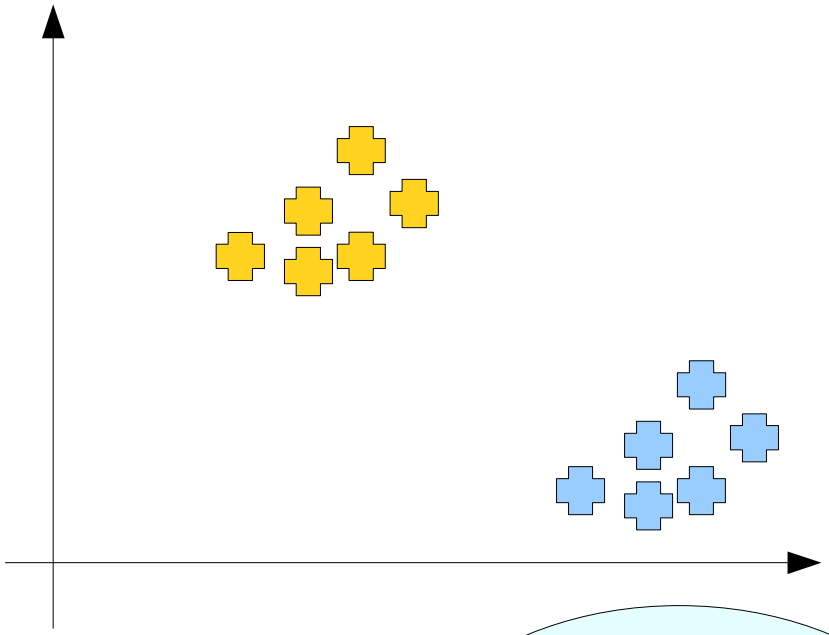


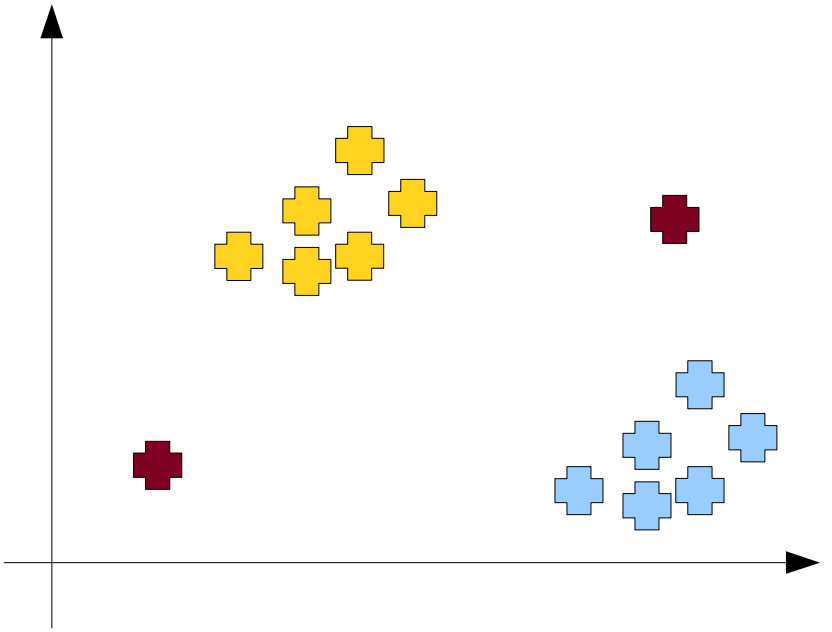
Step 3: Clustering

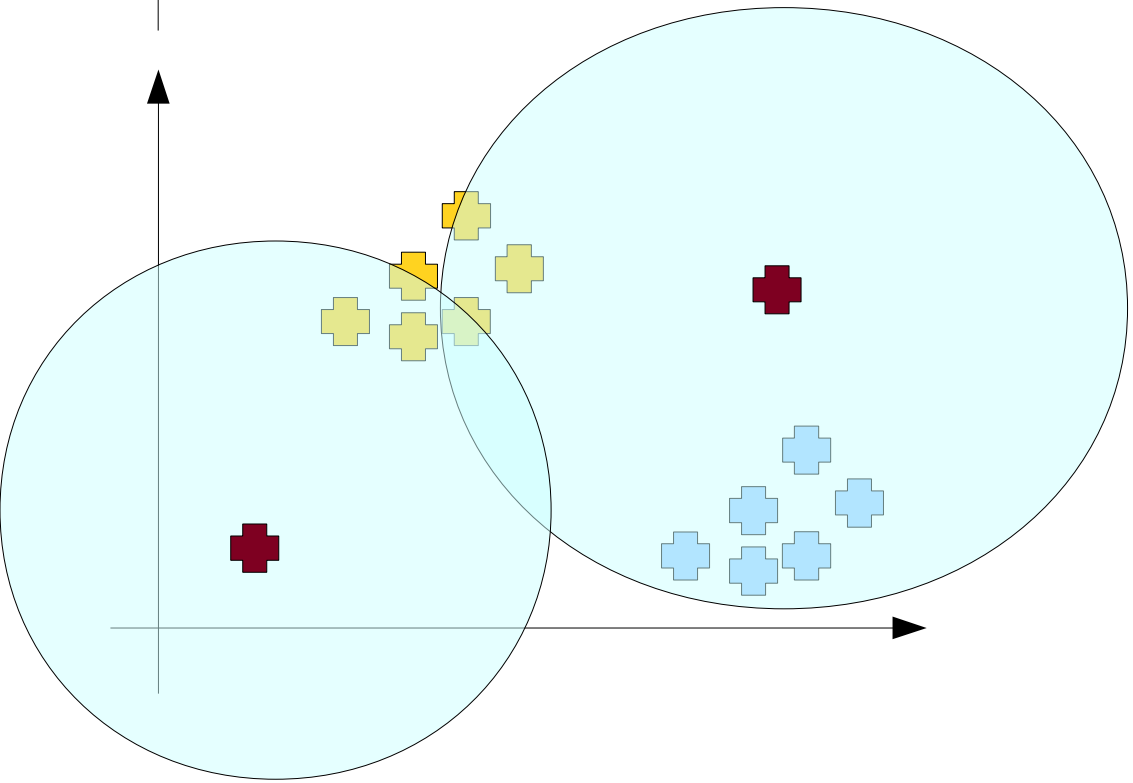


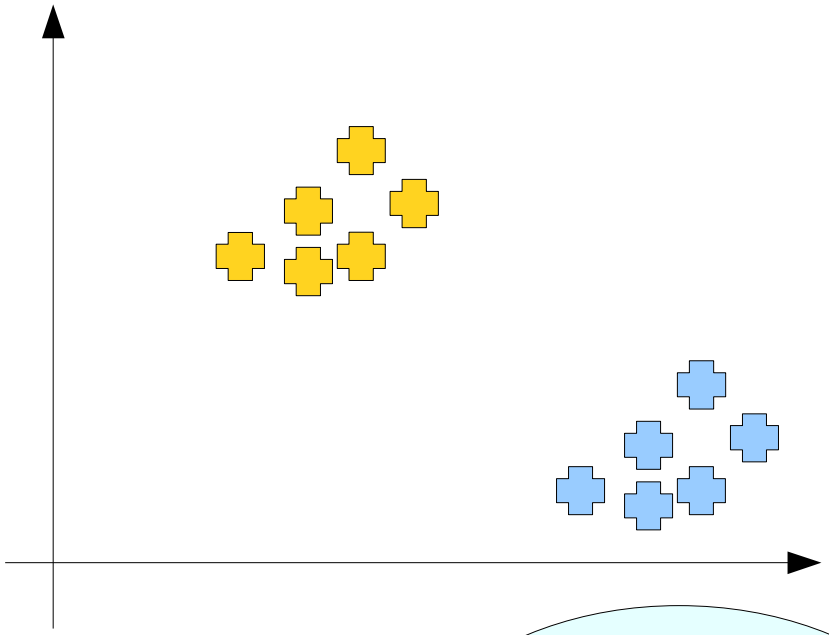


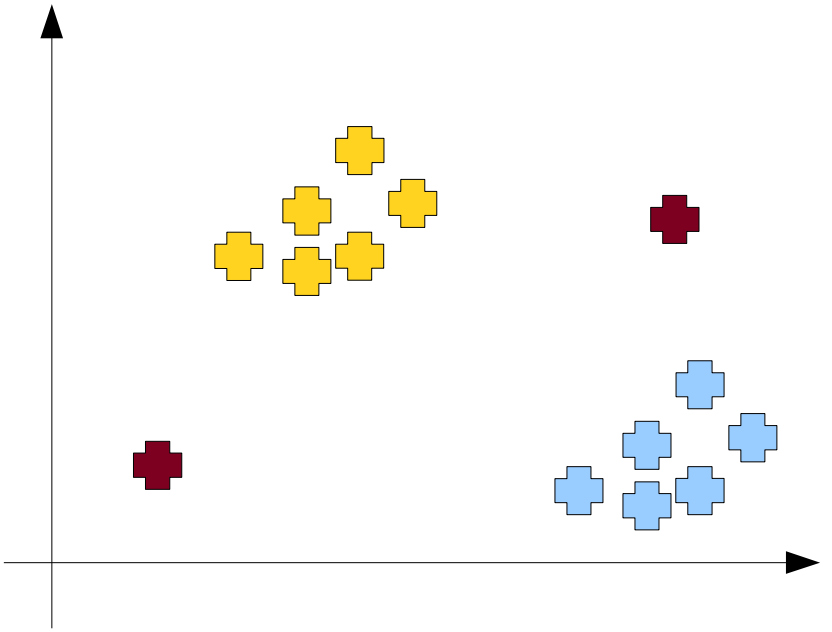


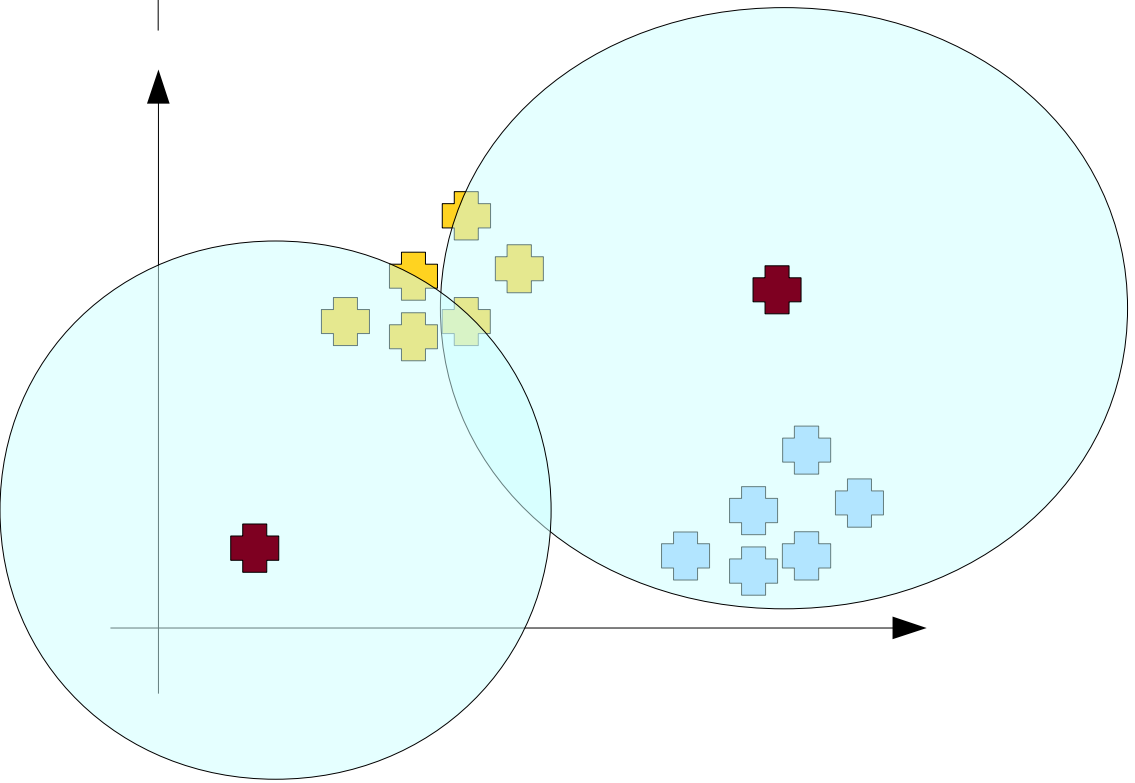


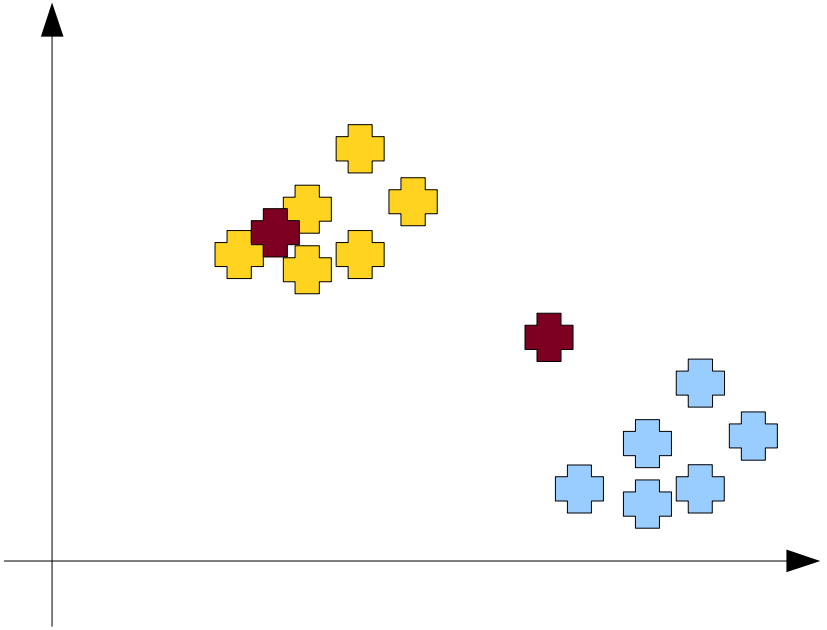


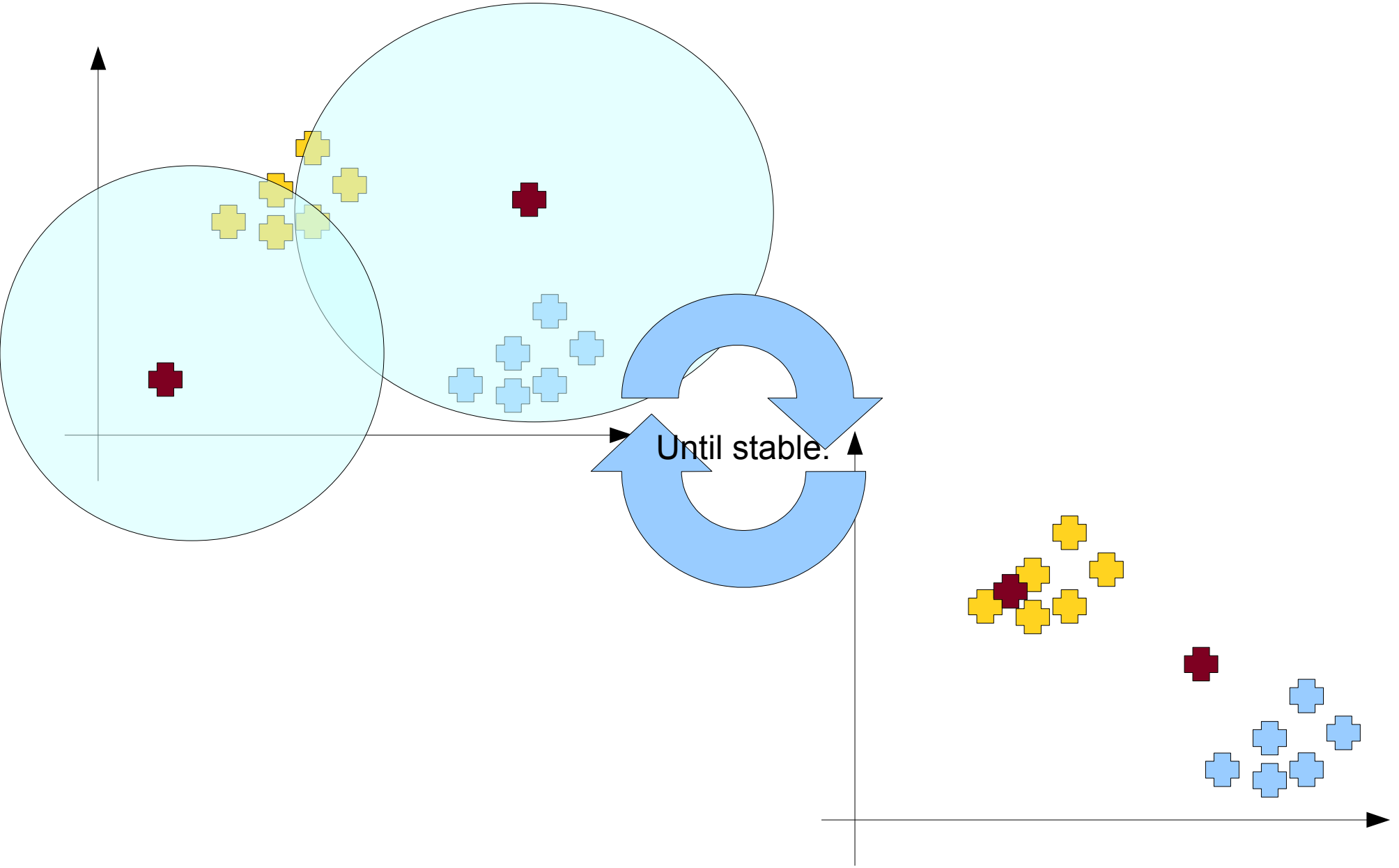











Until stable.

Reality

- Seed selection.
- Choice of initial k.
- Continuous updates.
- Regular addition of clusters.

Discover groups of similar items

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

From data to information.

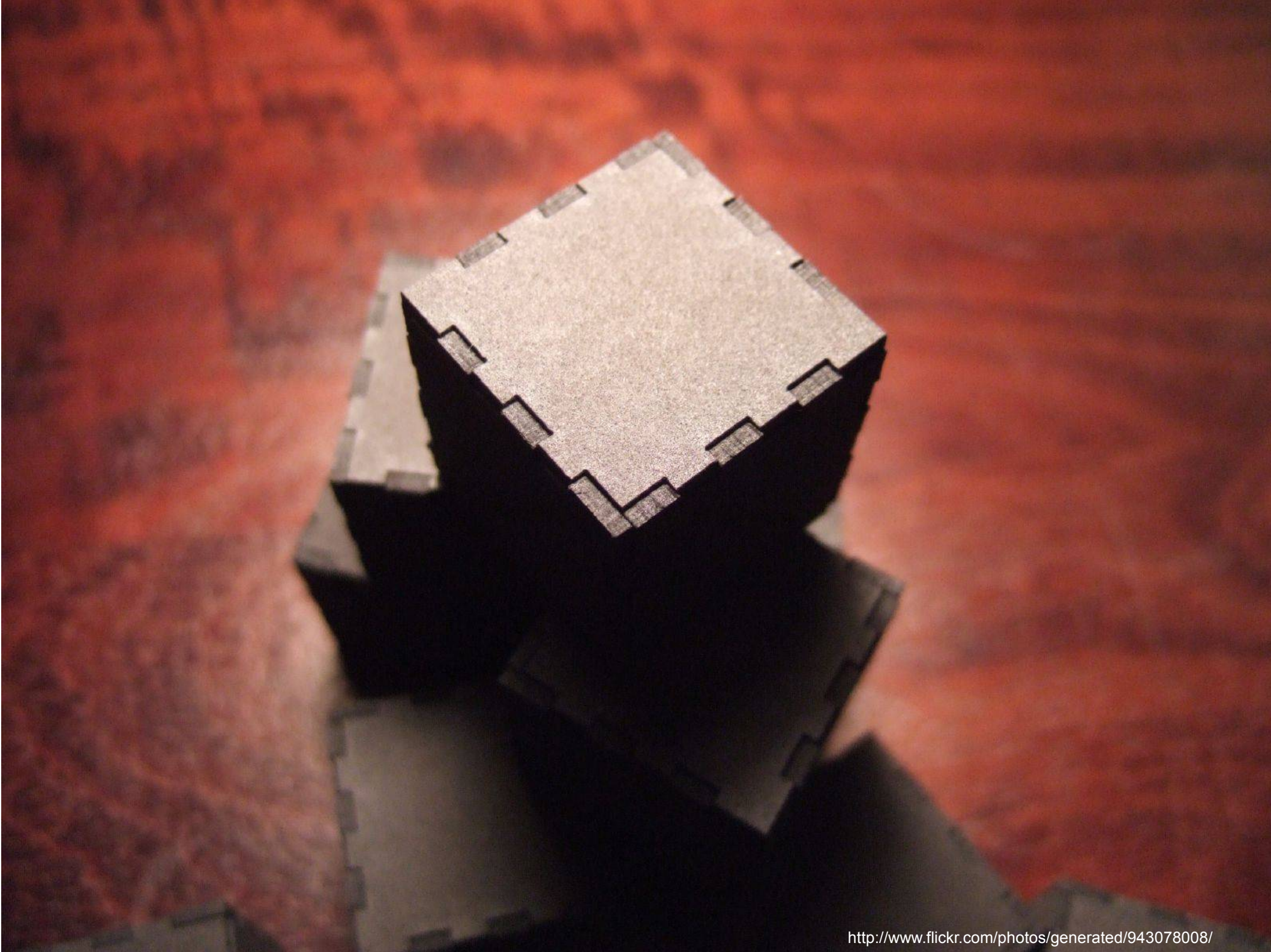
- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.

Evaluation

- Compare against gold standard.
- Use quality measures.
- Manual inspection.

From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.


http://www.flickr.com/photos/generated/943078008/


ttp://www.flickr.com/photos/eschipul/4160817135/

What else does Mahout have to offer.

Identify dominant topics

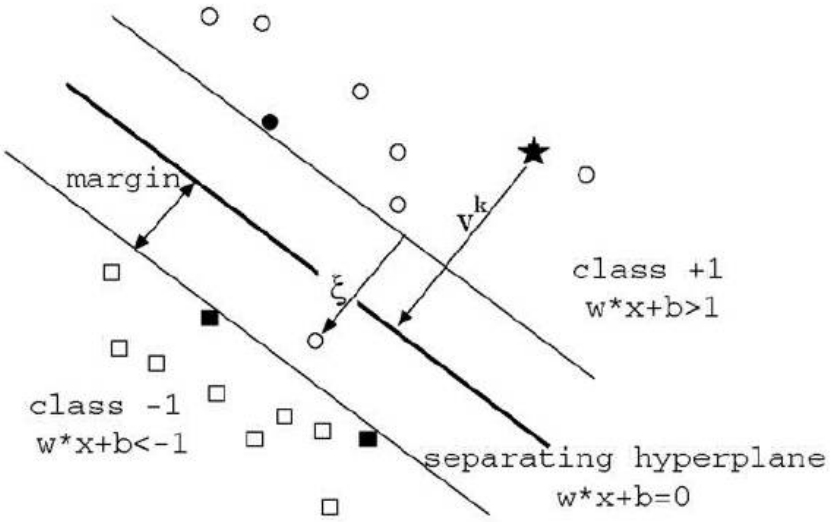
- Given a dataset of texts, identify main topics.

Algorithms: Parallel LDA

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

Assign items to defined categories.

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



By freezelight, http://www.flickr.com/photos/63056612@N00/155554663/

Google images oakland Search images Advanced Image Search

SafeSearch: Moderate ▼

Images Hide options Result

› **Any size**
Medium
Large
Icon
Larger than...
Exactly...

› **Any type**
Face
Photo
Clip art
Line drawing

› **Any color**
Full color
Black and white
Specific color

Related searches: oakland **raiders**

Oakland Airport
625 x 471 - 103k - jpg
visitingdc.com

Oakland Ranks Fifth
538 x 359 - 44k - jpg
bayareahomegirl.com

Oakland
900 x 600 - 171k - jpg
globalsecurity.org

OAKLAND First
400 x 400 - 27k
bayassociation.org

Oakland Gaudy Lexus 2
500 x 340 - 73k - jpg
lexusenthusiast.com

The **Oakland** Baseball
450 x 305 - 39k - jpg
sportsbusinesssims.com

Oakland
600 x 320 - 48k - jpg
webpages.scu.edu

Learn more about
550 x 366 - 56k - jpg
tripadvisor.com

Google images oakland Search images Advanced Image Search

SafeSearch: Moderate ▼

Images › Face Hide options

Results 1 - 20 of about 1,400,000 (0.

› Any size
Medium
Large
Icon
Larger than...
Exactly...

Any type
› Face
Photo
Clip art
Line drawing

› Any color
Full color
Black and white
Specific color

Reset options

Related searches: oakland raiders

Oakland, CA 94621
513 x 545 - 13k - gif
nflfootballstadiums.com

All Graphics »
262 x 278 - 43k - gif
coolchaser.com

Oakland Sideshow and
720 x 480 - 44k
channels.com

Detroit Tigers v
594 x 396 - 51k
zimbio.com

Oakland. by Davey D
359 x 512 - 26k - jpg
sfbayview.com

Detroit Tigers v
443 x 594 - 74k
zimbio.com

Dallas Cowboys v
594 x 404 - 60k
zimbio.com

Distributed by Tubemogul. The
720 x 480 - 18k
channels.com

oakland@coe.ufl.edu
379 x 471 - 110k - jpg
coe.ufl.edu

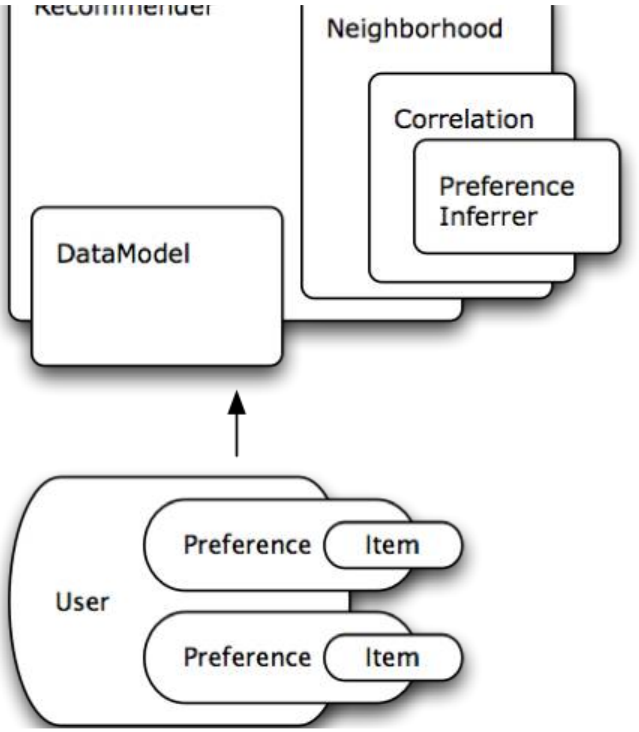
#20 of the **Oakland**
467 x 594 - 86k
zimbio.com

Assign items to defined categories.

- Naïve Bayes.
- Complementary naïve bayes.
- Random forests.
- Logistic regression.
- HMM for sequences.

Recommendation mining.

- Collaborative filtering.



Show most relevant ads

Show most relevant ads

clearasil"

erwandte Suchbegriffe: clearasil pickelstift, world of warcraft

este Ergebnisse « Zurück | Seit

Clearasil 44161 Tiefenreinigung Antibakterielle Reinigungspads, 60e

Neu kaufen: **EUR 5,99** 2 Angebote ab EUR 5,15

Lieferung bis **Samstag, 22. März**: Bestellen Sie innerhalb der nächsten **23 Stunden**
Kostenlose Lieferung möglich.

★★★★★ (1)

Drogerie & Bad: Alle 13 Artikel ansehen

Clearasil Ultra Anti-Pickel Reinigungspads, 65 Stück von Clearasil (**Ba**

Neu kaufen: **EUR 7,99**

Gewöhnlich versandfertig in 1 bis 3 Wochen.
Kostenlose Lieferung möglich.

Drogerie & Bad: Alle 13 Artikel ansehen

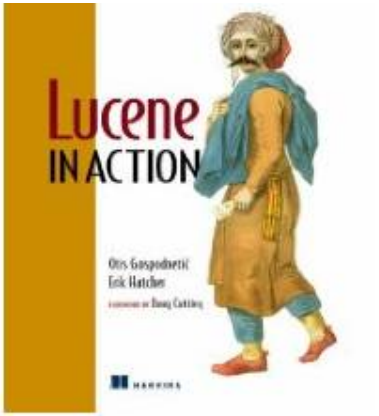
World of WarCraft: Wrath of the Lich King (Add-on) von Vivendi Unive
Vista / XP)

Neu kaufen: **EUR 39,99**

Vorbestellbar
Kostenlose Lieferung möglich.

Games: Alle Artikel ansehen

Show most relevant ads



See larger image

Share your own customer images

Publisher: learn how customers can search inside this book.

~ Otis Gospodnetic (Author), Erik Hatcher (Author)
(20 customer reviews)

List Price: $44.95
Price: $29.67 & this item ships for **FREE with Super Saver Shipping**. Details
You Save: $15.28 (34%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

23 new from $19.99 **15 used** from $17.11

Also Available in:	List Price:	Our Price:	Other Offers:
Paperback (2)	$44.99	$29.69	

Frequently Bought Together

Customers buy this book with Building Search Applications: Lucene, LingPipe, and Gate by Manu Konchady

+

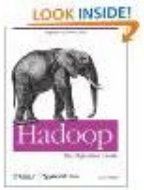
Price For Both: $70.12

Add both to Cart | Add both to Wish List

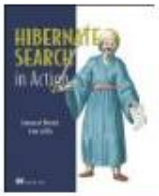
Show availability and shipping details

Customers Who Bought This Item Also Bought

Solr 1.4 Enterprise Search Server by David

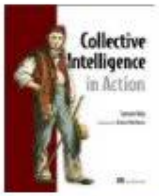


Hadoop: The Definitive Guide by Tom White



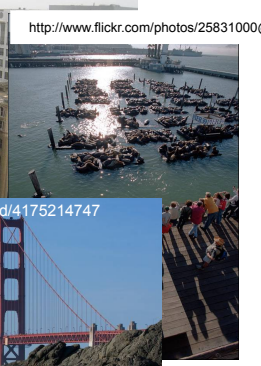
Hibernate Search in Action by Emmanuel

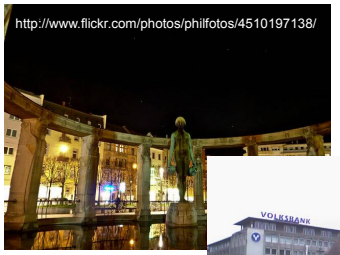
Introduction to Information Retrieval by

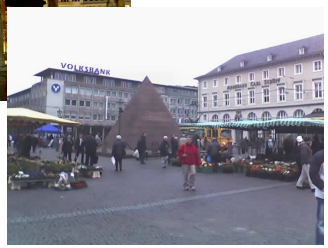


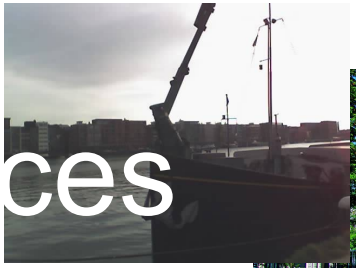
Collective Intelligence in Action by Satnam Alag

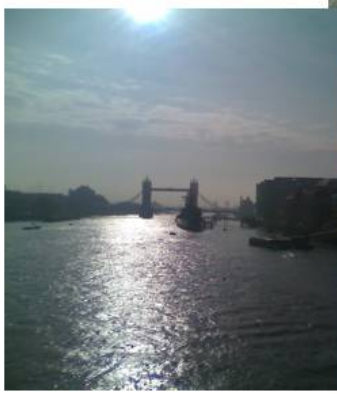
Recommending places

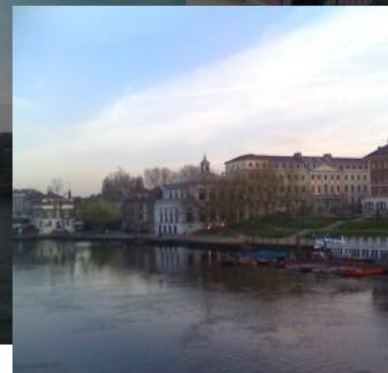












Thanks to Falko Menge for the pictures of Brussels.

Recommending people

Recommendation mining.

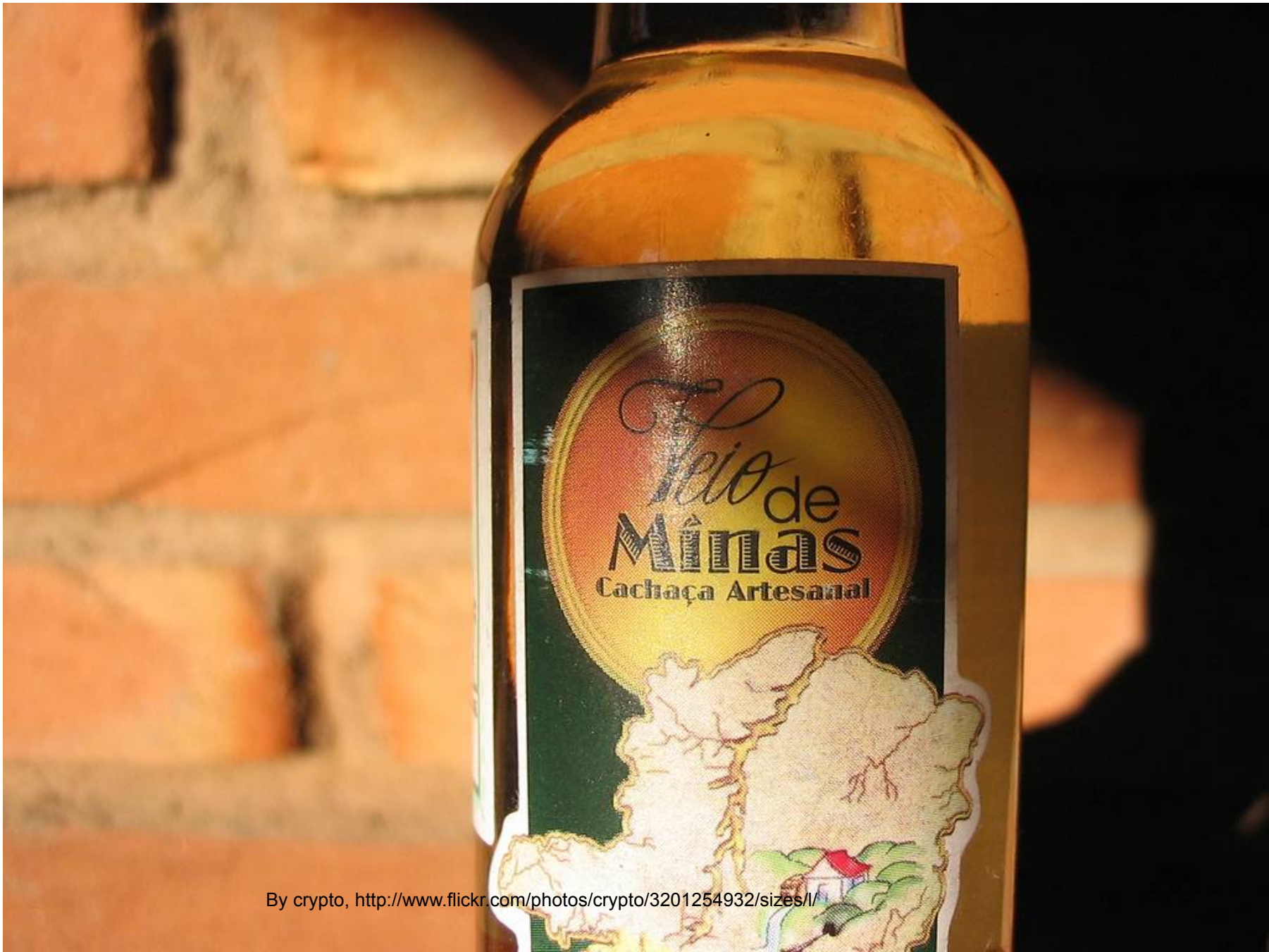
- Online collaborative filtering on single machine.
- Offline Map/Reduce based version.
- Content similarity can be integrated.

- Based on former Taste project.

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.



By crypto, http://www.flickr.com/photos/crypto/3201254932/sizes/l/

By libraryman, http://www.flickr.com/photos/libraryman/78337046/sizes/l/

rypto/3201254932/sizes/l/

By quinnanya, http://www.flickr.com/photos/quinnanya/2806883231/

By crypto, http://www.flickr.com/photos/crypto/3201254932/sizes/l/

By libraryman, http://www.flickr.com/photos/libraryman/78337046/sizes/l/

Requirements to get started

March 14, 2009 by Artful Magpie
http://www.flickr.com/photos/kmtucker/3355551036/

amazon web services™

Sign in to the AWS Management Console | Create an AWS Account

AWS | Products | Developers | Community | Support | Account

Products & Services

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2 Details

- **EC2 Overview**
- FAQs
- Amazon EC2 SLA
- EC2 Instance Types

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain

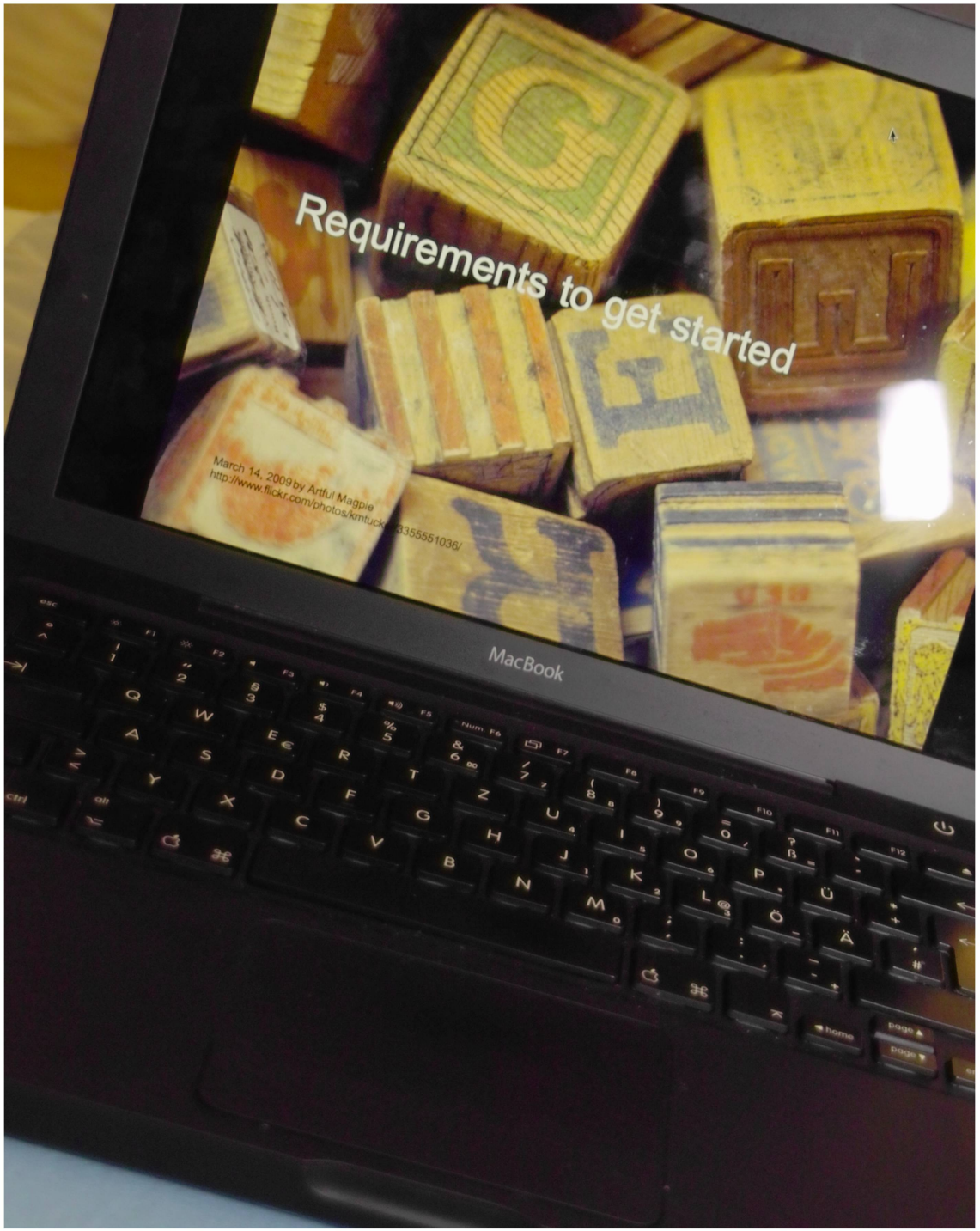
Sign Up For Amazon EC2

Amazon Elastic MapReduce

Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).

Using Amazon Elastic MapReduce, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial

(Thanks to Thilo for helping set up the cluster, Thanks to packet and masq for two of the three machines.)


Requirements to get started
March 14, 2009 by Artful Magpie
http://www.flickr.com/photos/kmfuck
3355551036/
MacBook

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 a committer.


Sebastian Schelter
Jake Mannix
Benson Margulies
Robin Anil
David Hall
AbdelHakim Deneche
Karl Wettin
Sean Owen
Grant Ingersoll
Otis Gospodnetic
Drew Farris
Jeff Eastman
Ted Dunning
Isabel Drost

Become a committer: Of Apache Mahout

Emeritus:

Niranjan Balasubramanian
Erik Hatcher
Ozgur Yilmazel
Dawid Weiss

-user@.apache.org

-dev@.apache.org

Image by: Patrick McEvoy

Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.


Herzlich Willkommen!
Thanks to Tim Lossen et. al for taking amazing pictures of the conf.

Berlin Buzzwords 2011

Search/ Store/ Scale

May/ June 2011

Thanks to Tim Lossen et. al for taking amazing pictures of the conf.

*-user@hadoop.apache.org

*-dev@hadoop.apache.org

Image by: Patrick McEvoy

Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.