

Apache Mahout

Bringing Machine Learning to Industrial Strength

Agenda

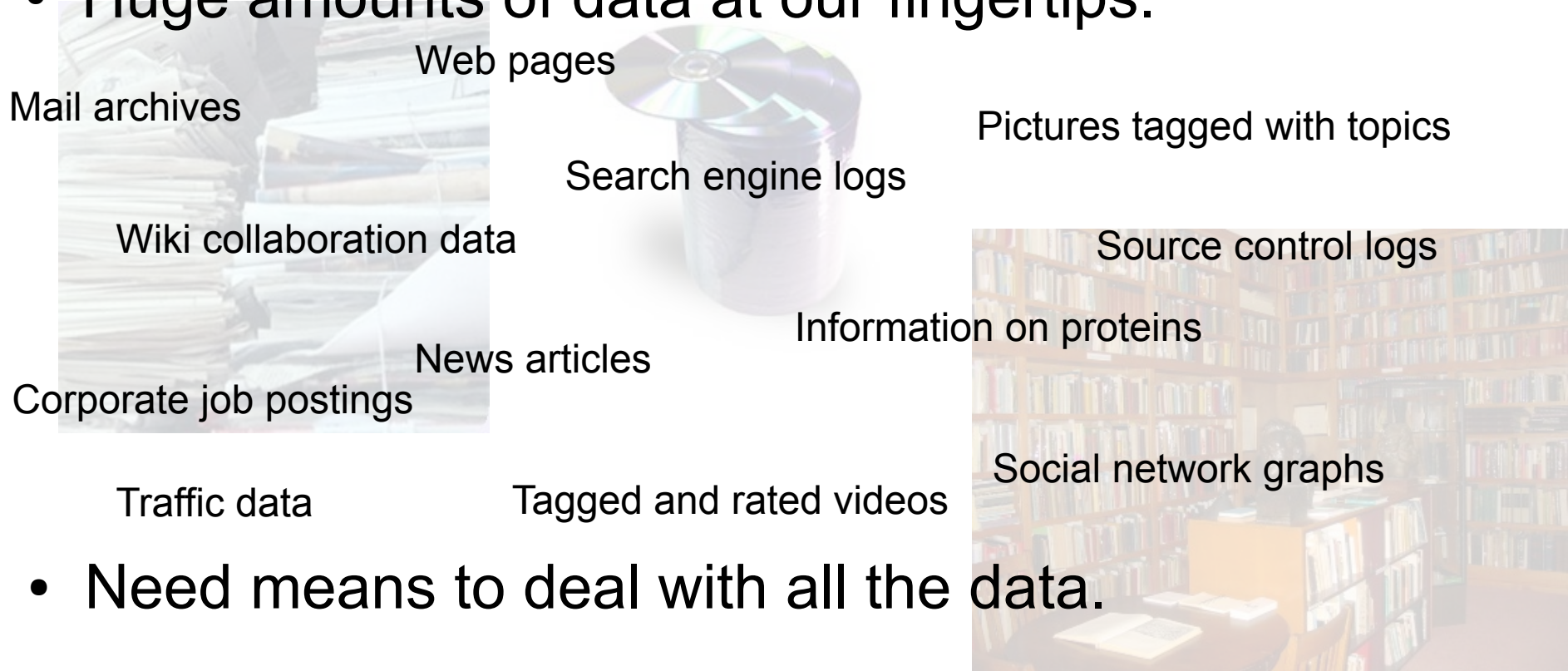


- What is machine learning all about?
 - The mission of Mahout.
 - Let's get to work:
 - Grouping data into topics.
 - Assigning data to pre defined categories.
 - Recommend items to users.
 - Example applications!
-
-

Problem setting



- Huge amounts of data at our fingertips.

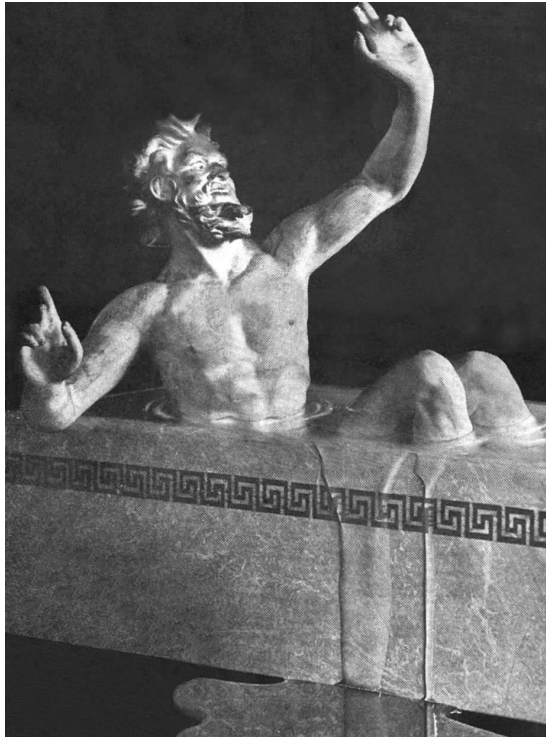


- Need means to deal with all the data.
-

Problem setting



- Nature generates data.
- Archimedes generates model.



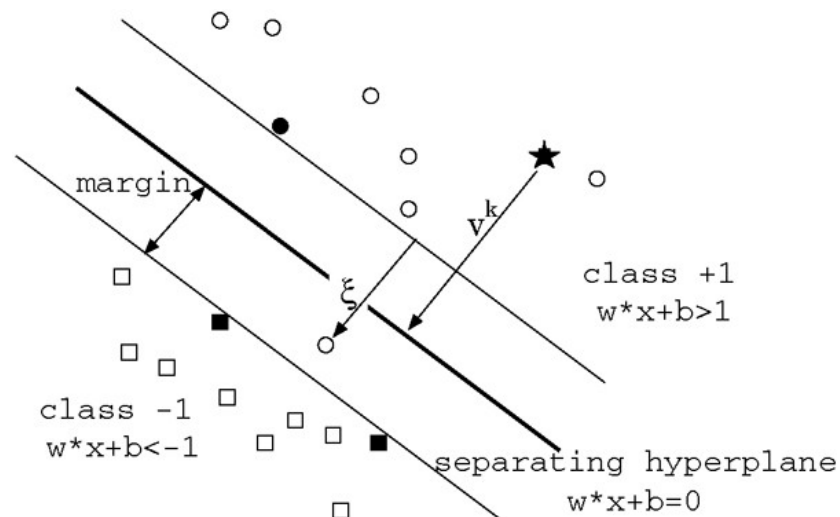
$$\frac{\text{Density of Object}}{\text{Density of Fluid}} = .$$

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

Problem setting



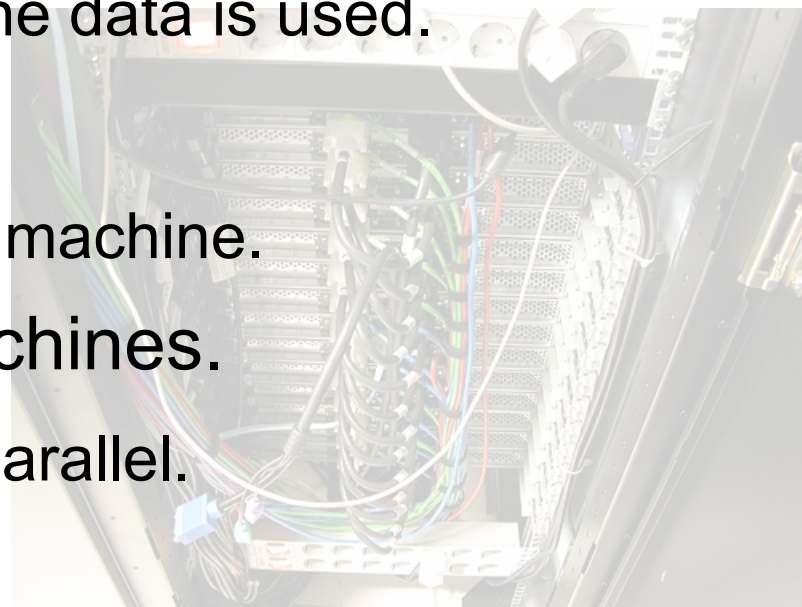
- Nature generates data.
- ML generates models.



Dataset sizes.



- Dataset usually are huge.
- Solution one: Use only a sample of the data.
 - Problem: Not all information in the data is used.
- Solution two: Use all the data.
 - Problem: Takes too long on one machine.
- Our Solution: Use multiple machines.
 - Handle all data, but process in parallel.



Our mission



- Build data (text) mining algorithms that are scalable.
- Context:



Hadoop – one way of parallelizing algorithms

Once upon a time



- How it all began:
 - Summer 2007: Crazy developers needed scalable ML.
 - Mailing list and wiki followed quickly.
 - Rather large community even before project start.
 - 25.01.2008: Project Mahout launched.
 - Today: Mahout @ FrOSCon.
-
-

Who we are



Dawid Weiss
Carrot2



Karl Wettin
Lucene



Grant Ingersoll
Lucene PMC



Otis Gospodnetic
Lucene



Jeff Eastman
Welcome!



Ted Dunning
The Veoh guy

Erik Hatcher
Lucene

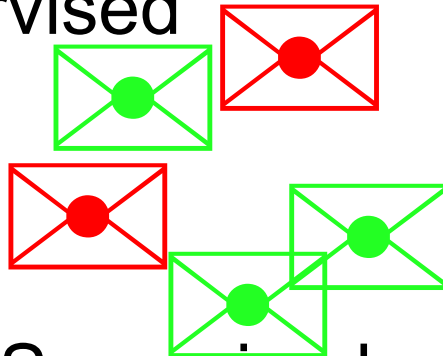


Isabel Drost
(that would be myself)

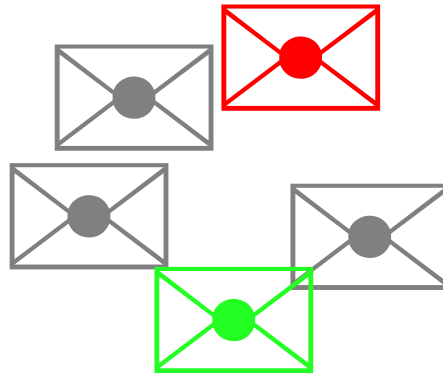
Types of learning tasks



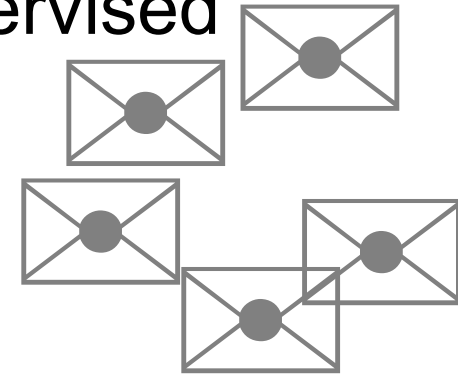
- Supervised



- Semi Supervised



- Unsupervised



Types of learning tasks



- Supervised

e.g. Classification

- Semi Supervised

- Unsupervised

e.g. Clustering

Template for learning



- Get the data.
 - Transform data to machine understandable form.
 - Choose an appropriate algorithm based on problem.
 - Search for best parameters.
 - Combine features, found parameters, and algorithm.
-
-

Template for learning



- Get the data.
 - Transform data to machine understandable form.
Features
 - Choose an appropriate algorithm based on problem.
No single best
 - Search for best parameters.
Evaluate.
 - Combine features, found parameters, and algorithm.
-
-

Clustering



- Example problem setting:
 - What you have: Huge amount of mails, say Debian lists.
 - What you want: Mails grouped by common topic.
- Algorithms so far: K-Means, Canopy, Mean Shift, Hierarchical.



Clustering

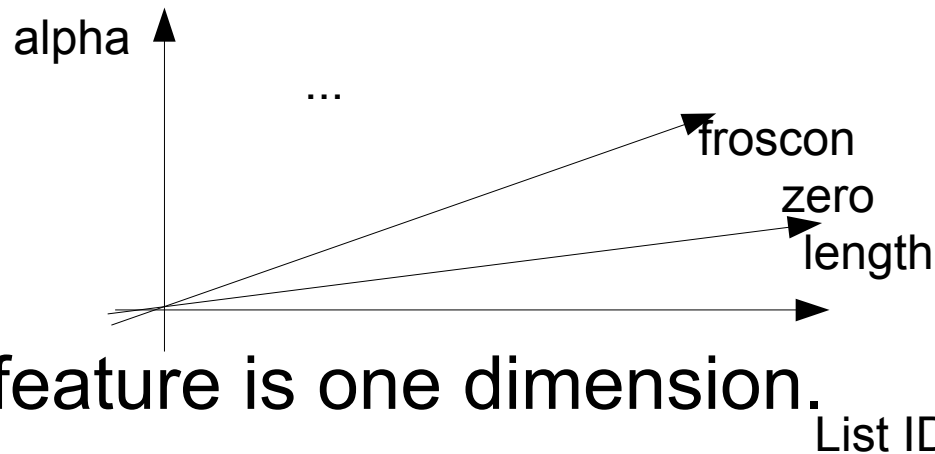


- 1) Gather the mails (e.g. from the archives).
 - 2) Find a way to generate vectors of mail properties.
 - Mailinglist ID.
 - Length of the mail.
 - Parse the text and make features from word occurrence.
 - Parse the subject line and make features from words.
 - ...
 - 3) Apply some clustering algorithm to data vectors.
-

Clustering – step 2



- Each mail is a point in high dimensional space.



One of your mails:

$$\begin{pmatrix} 0 \\ 1 \\ 10 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

- Each feature is one dimension.
- k-Means: Group points that are close to each other.
- Meaning of “close” depends on use case.

Clustering – step 2



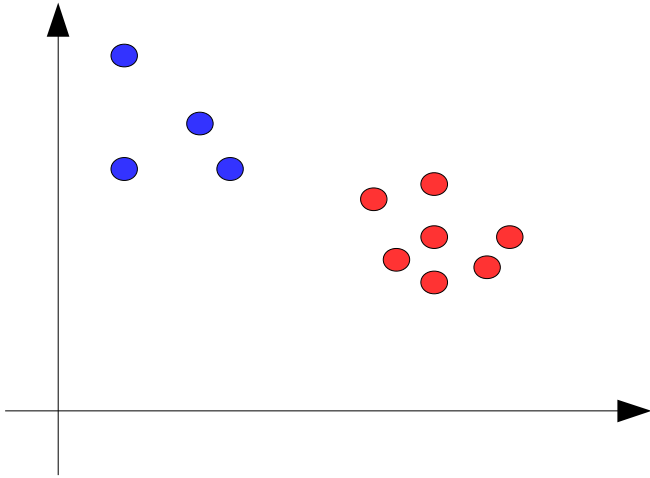
- Common text feature generation schemes:
 - One binary dimension per term.
 - TF – Each dimension counts term occurrences.
 - TFIDF – Weighted number of terms.
 - Specific features depend on your application.
-
-

Clustering: k-Means – step 3

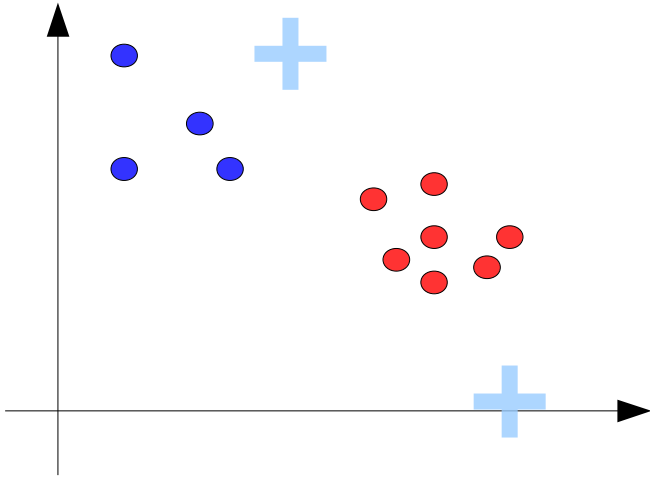


- Choose how many clusters/topics you seek (k).
 - Choose arbitrary centers.
 - Choose your definition of “closeness”.
 - Repeat:
 - Assign each point to its closest the center.
 - Use the mean of the point groups as new center.
 - Until assignment stable.
-
-

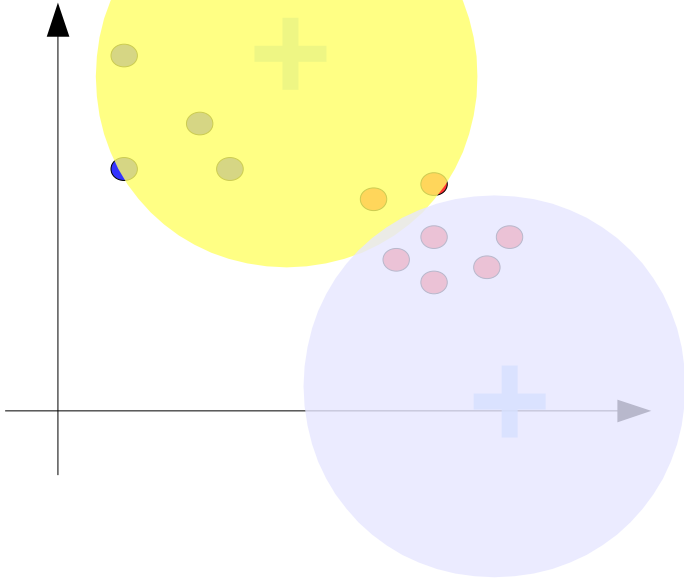
Clustering: k-Means – step 3



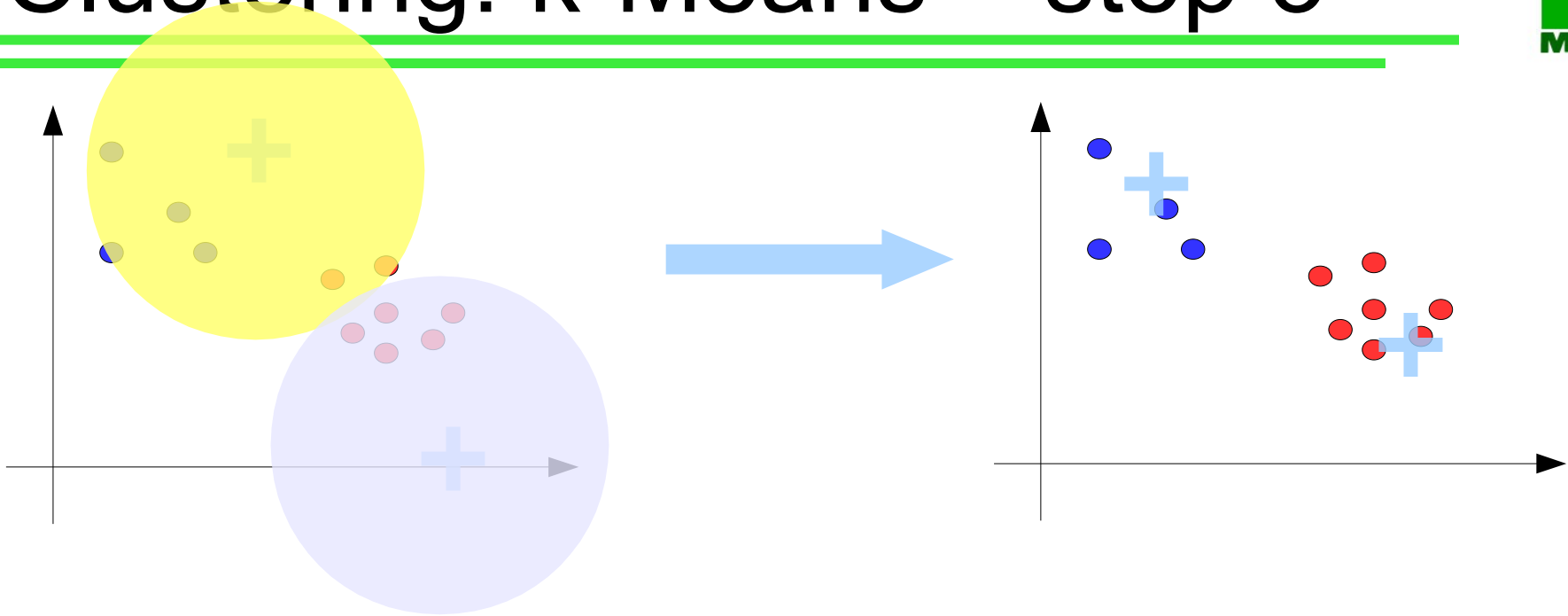
Clustering: k-Means – step 3



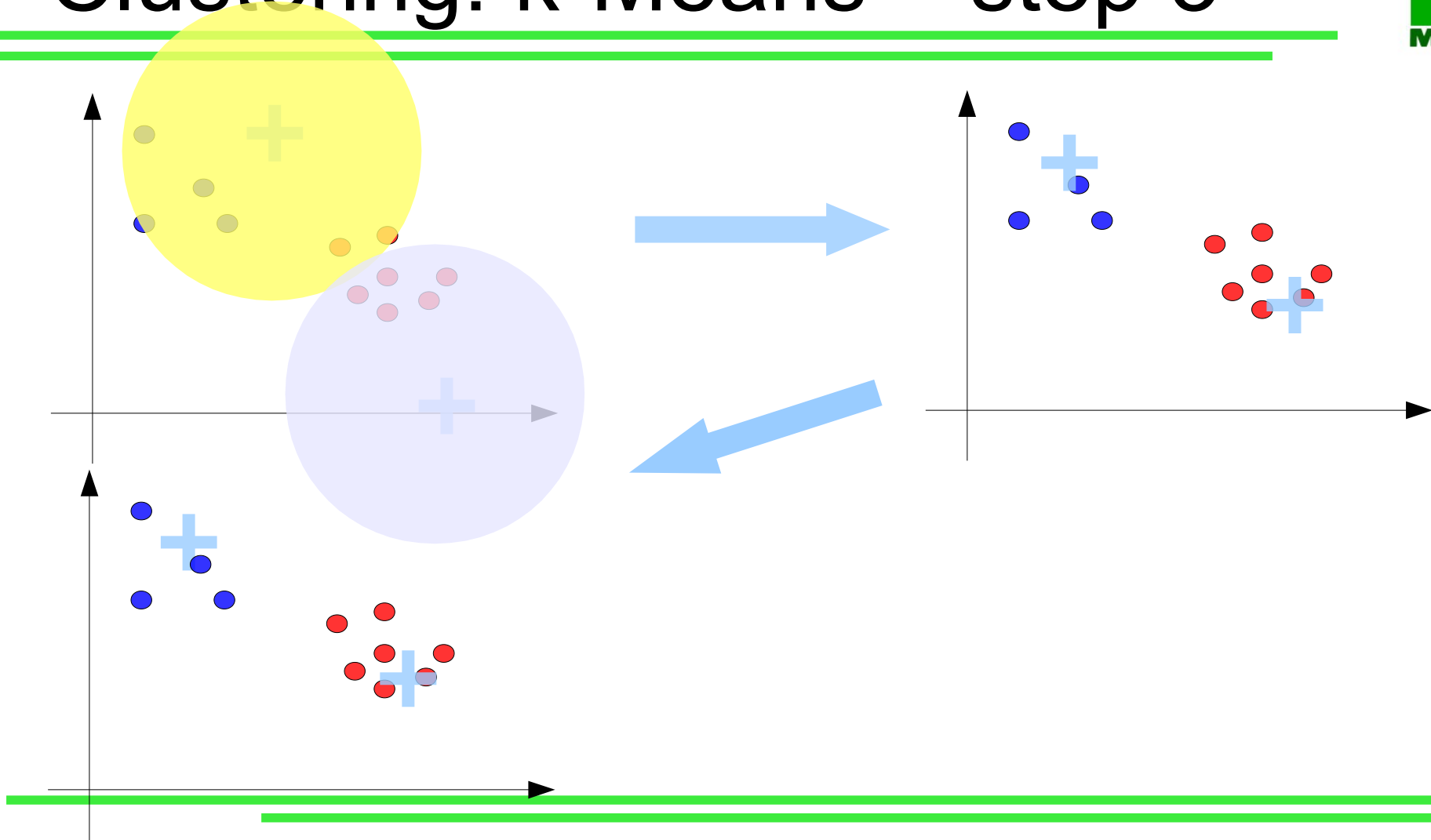
Clustering: k-Means – step 3



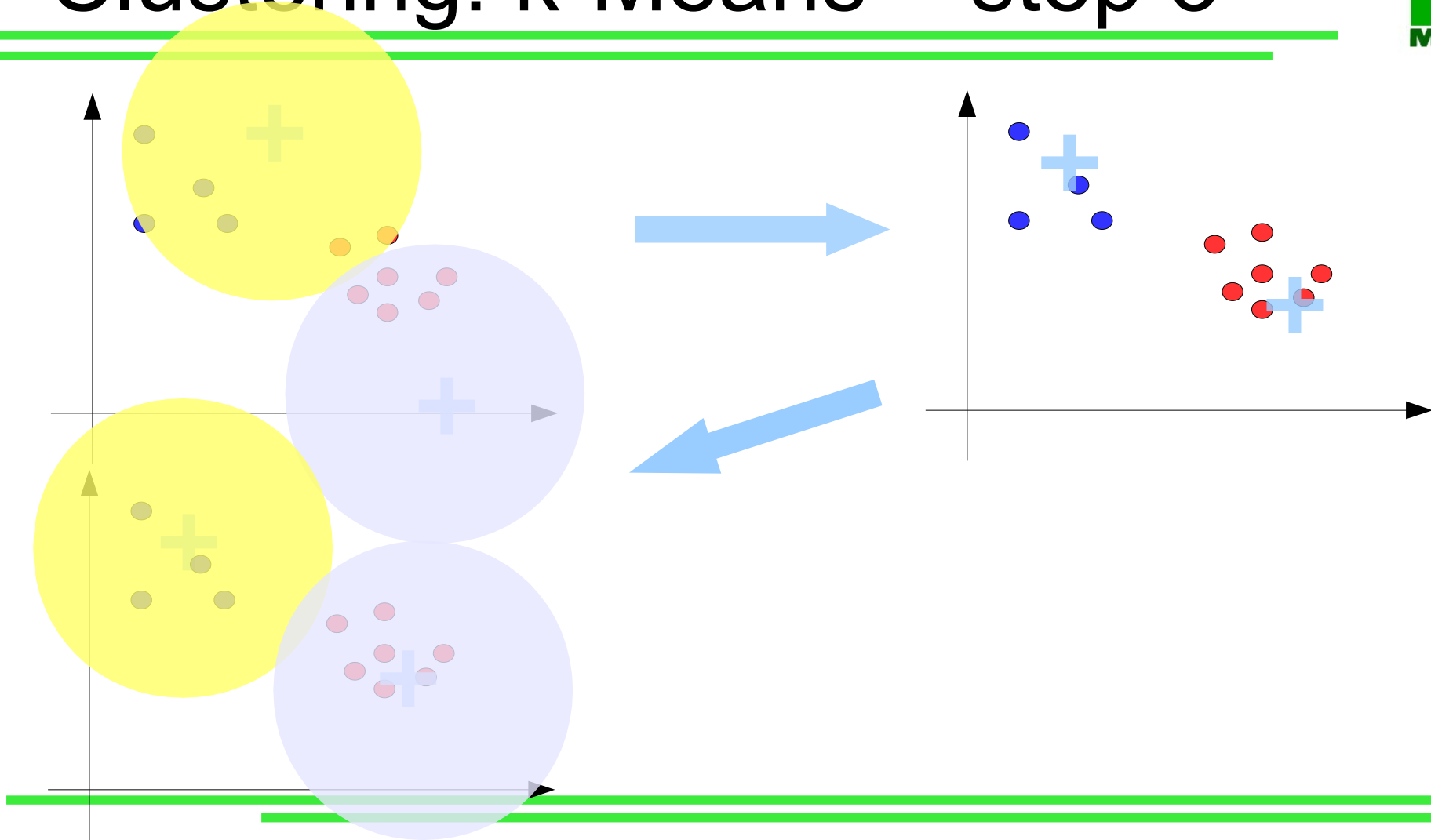
Clustering: k-Means – step 3



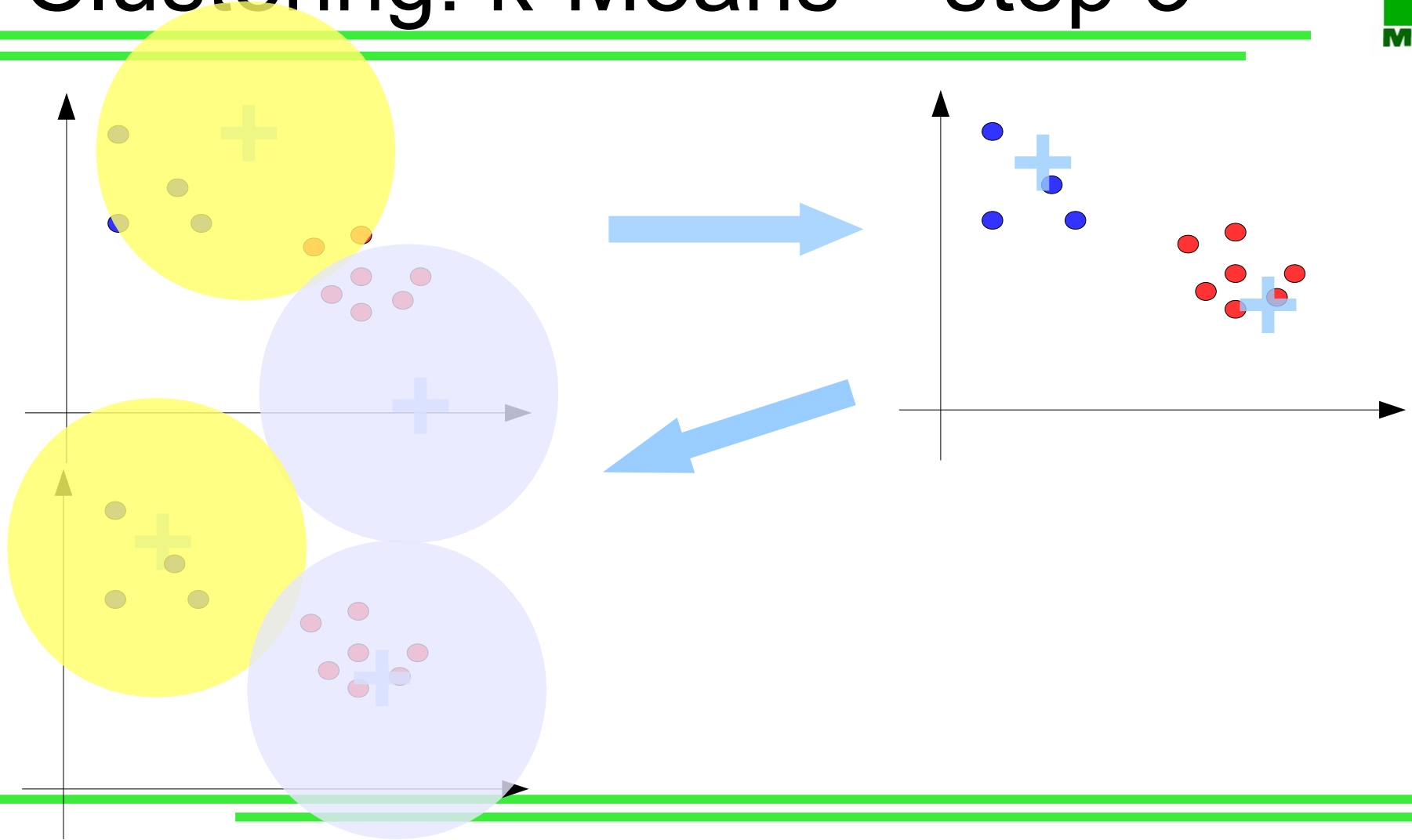
Clustering: k-Means – step 3



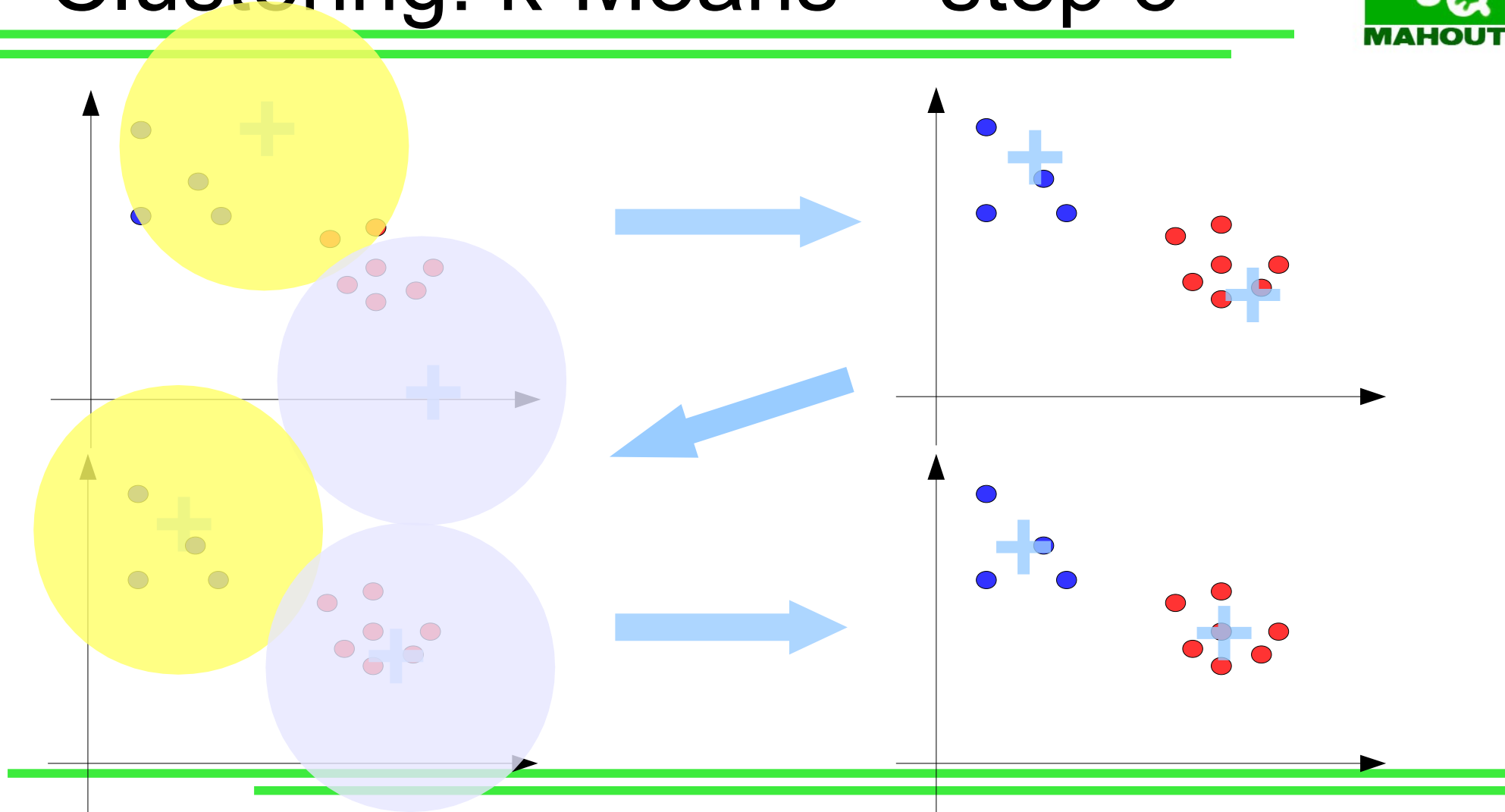
Clustering: k-Means – step 3



Clustering: k-Means – step 3



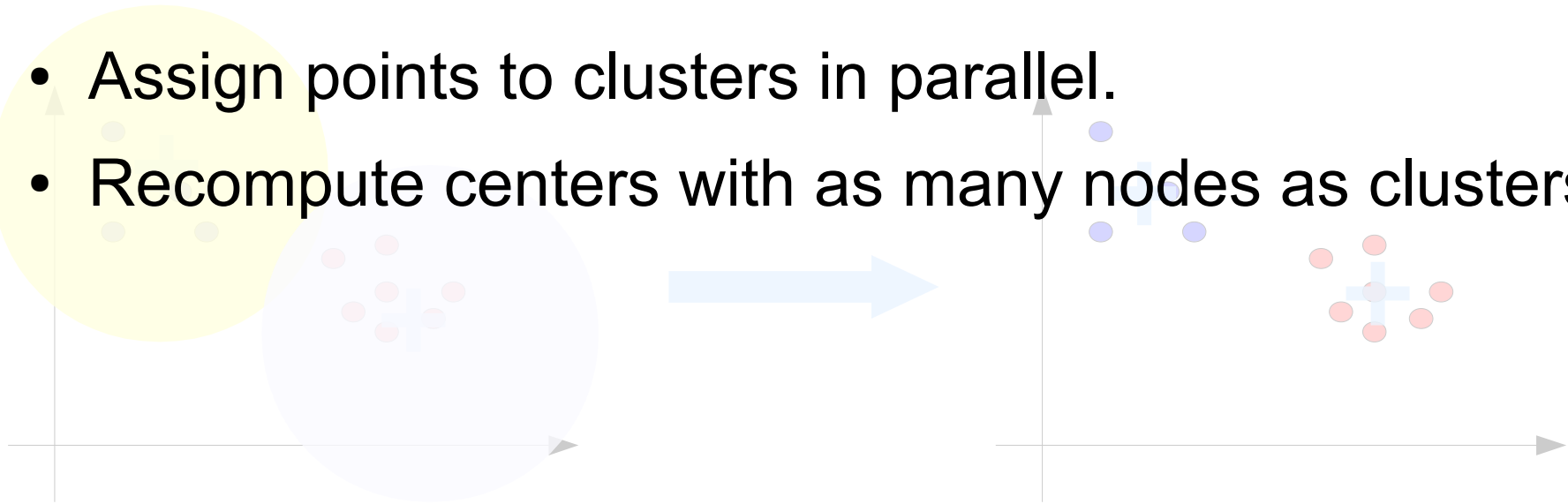
Clustering: k-Means – step 3



Clustering: k-Means – step 3



- Distributed version:
- Assign points to clusters in parallel.
- Recompute centers with as many nodes as clusters.



Points for optimization



- Representation of objects as feature vectors.
 - Definition of “distance” of vectors.
 - Definition of cluster center.
 - Starting points.
 - k-Means trivial but not necessarily best algorithm.
 - Evaluation: Usually against gold standard.
-
-

Points for optimization

- Representation of objects as feature vectors. Features
 - Definition of “distance” of vectors.
 - Definition of cluster center. Parameters.
 - Starting points.
 - k-Means trivial but not necessarily best algorithm. No single best.
 - Evaluation: Usually against gold standard.
-
-

Classification



- Example problem setting:
 - You have: Set up a search engine.
 - You want: To index pages of your favorite topic.
- Algorithms planned: Naïve Bayes, logistic regression, SVM.

[22C3: Private Investigations](#)

The 22nd Chaos Communication Congress (**22C3**) is a four-day conference. More information on **22C3** is coming up. For now, study and distribute our CC-BY-SA licensed content. www.ccc.de/congress/2005/ - 3k - [Cached](#) - [Similar pages](#)

[CCC | Chaos Communication Congress](#) - [[Translate this page](#)]

22C3: Private Investigations. Der 22. Chaos Communication Congress bis 30. ... All Informationen zum **22C3** finden sich unter [Internet Link] www.ccc.de/congress/ - 13k - [Cached](#) - [Similar pages](#)
[[More results from www.ccc.de](#)]

[Upcoming.org: 22C3: Private Investigations at Berliner Congress ..](#)

The 22nd Chaos Communication Congress (**22C3**) is a conference on technical and social aspects of the Internet. The 22nd Chaos Communication Congress **22C3**: Private Investigations. upcoming.org/event/23579/ - 9k - [Cached](#) - [Similar pages](#)

[22c3 h07 - NerdPedia](#)

Darum: META-REFRESH <http://hackerhippie.org/newiki/22c3>. [old content due to feature of this wiki] ... www.nerdpedia.org/index.php/22c3_h07 - 8k - [Cached](#) - [Similar pages](#)

[22C3 \(Private Investigations\) - Call for Papers | Uwe Hermann](#)

The first signs of the upcoming **22C3** congress ("Private Investigations") are here. The 22nd Chaos Communication Congress (**22C3**) is a four-day conference. www.hermann-uwe.de/blog/22c3-private-investigations-call-for-papers - 44k

[The Lunatic Fringe » Blog Archive » 22C3 Updates](#)

22C3 Updates. Sunday October 23rd 2005, 18:41 Filed under: General. The months have been quite exhausting in terms of getting people on track for the congress. tim.geekheim.de/2005/10/23/22c3-updates/ - 17k - [Cached](#) - [Similar pages](#)

[22C3 \(Plan 9 wiki\)](#)

The 22nd Chaos Communication Congress Where: Berlin, Germany When: 27-30. October 2005 www.ccc.de/congress/. Who will be there? 20h, MTG, garbeam, uwe. www.ccc.de/wiki/plan9/22C3/ - 2k - [Cached](#) - [Similar pages](#)

[maha's blog » 22c3](#)

22c3. Sunday, July 10th, 2005. Gestern fand die erste Sitzung der ... You are currently browsing the archives for the **22c3** category. ... www.maha-online.de/blog/category/ccc/22c3/ - 14k - [Cached](#) - [Similar pages](#)

[symlink.ch | 22C3: Fnord-Jahresrückblick oder nicht? - \[Translate t](#)

22C3: Fnord-Jahresrückblick oder nicht?' | Einloggen/Account erstellen | 9 Kommentare ... Re: Bin gespannt auf **22c3** by Anonymer Feigling Wednesday

Classification

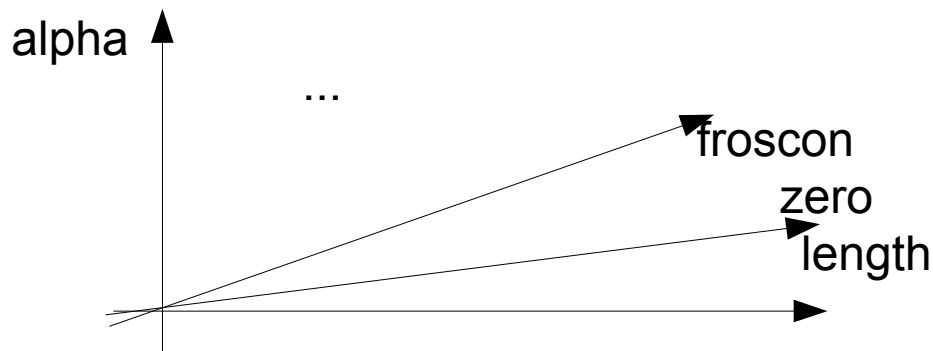


- 1) Gather web pages.
 - 2) Manually assign labels “off-topic” and “on-topic”.
 - 3) Generate vectors of web page properties.
 - Parse the text and make features from word occurrence.
 - Length of web page.
 - 4) Train some classification algorithm to labeled data.
 - 5) Apply the trained algorithm to new incoming data.
-
-

Classification



- Each web page is point in high dimensional space:



One of your mails:

$$\begin{pmatrix} \cdot \\ 1 \\ 10 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

- Classifier learns properties relevant for your topic.
- Which properties are selected depends on labels.

Classification: Naïve Bayes



- Sequential version:
 - For each label count terms per class.
 - Create model that represents label and feature counts.

$$p(C|F_1, \dots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^n p(F_i|C)$$

- Distributed version:
 - Counting feature occurrences for each label.
-
-

Points for optimization



- Representation of objects as feature vectors.
- Selection of labeled training data.
- Naïve Bayes easy, but not necessarily best:

SVM

Logistic Regression

Perceptron

Winnow

...

Rocchio

Points for optimization

- Representation of objects as feature vectors.
- Selection of labeled training data.
- Naïve Bayes easy, but not necessarily best:

SVM

Logistic Regression

Perceptron

No single best algorithm.

Winnow

Rocchio

...

Recommendation Mining



- Example problem setting:

- You have: A video sharing web service + terabytes of log files of user interactions.
- You want: Recommend users videos they might like.

- Integrated: Taste



Recommendation Mining



- 1) Gather user interaction logs.
 - For each user store which videos the user watched.
 - For each video store additional information (year, actors...)
 - 2) Create feature vectors from additional information.
 - 3) Recommend videos based on:
 - Similarity to the videos the user watched.
 - Videos other users with same preferences watched.
-
-

Points for optimization



- Use video ratings as additional information.
 - Find additional video information.
 - Use implicit user feedback:
 - How long/often did the user watch the video?
 - Did the user recommend the video to others?
 - Definition of user similarity is up to you.
 - Definition of video similarity is up to you.
-
-

Points for optimization



- Use video ratings as additional information.
- Find additional video information.
- Use implicit user feedback:
 - How long/often did the user watch the video?
 - Did the user recommend the video to others?
- Definition of user similarity is up to you.
- Definition of video similarity is up to you.

Initial Mahout goals



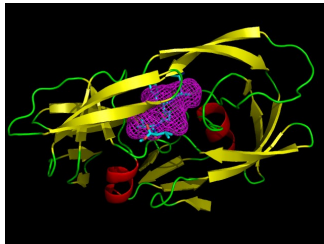
- **Clustering** (k-Means, Expectation Maximization, Mean Shift, Canopy, Hierarchical Clustering).
 - **Classification** (Naïve Bayes, Logistic Regression, Support Vector Machines).
 - **Recommendation mining** (Taste).
 - **Regression** (Linear Regression).
 - **Dimensionality reduction** (Principal Components Analysis, Independent Components Analysis, Gaussian Discriminative Analysis).
-
-

GSoC @ Mahout



- Genetic Algorithms for Mahout
 - Several ways to parallelize.
 - Good for complex problems.
 - Naïve Bayes and Complementary Naïve Bayes.
 - Classification algorithm.
 - Scale well to large amounts of data.
 - Straight forward to parallelize.
 - Two advisors: Mahout + University.
-
-

Example Applications



- Top Stories
- World
- U.S.
- Business
- Elections
- Sci/Tech
- Entertainment
- Sports
- Olympics
- Health
- Most Popular
- News Alerts
- RSS | Atom
- About Feeds
- Mobile News
- About
- Google News



China Daily

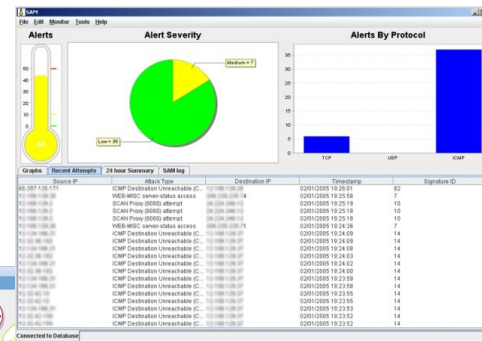
Olympics
Johnson ends gold drought with Olympic win in balance beam
Kansas City Star - **51 minutes ago**
US gymnast Shawn Johnson had a big smile and a gold medal at the end of Tuesday's balance beam. BEIJING | Jerry Seinfeld always says that the silver medal at the Olympics has to be the worst one.
Video: Olympics 08: Johnson, Liukin Take More Medals AssociatedPress
Iowa's silver belle spins gold Minneapolis Star Tribune
Seattle Times - The Associated Press - USA Today - Kansas City Star
all 2,969 news articles »



Canoe.ca

The Making Of China's Olympic Golden Age
CBS News - **43 minutes ago**
China is obsessed with Olympic gold that it is training 200,000 handicapped kids in state-run sports boarding schools. (CBS) CBS News staffers file insider impressions and share their experiences throughout the day.
Americans' necks full of Olympic medals, but precious few are gold Dallas Morning News
Chinatown gets a lift from Beijing Olympics San Francisco Chronicle

News



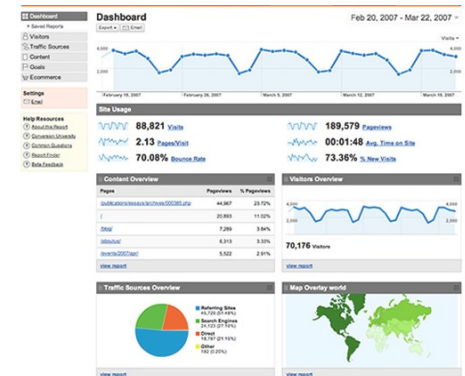
The Apache
The Powerful #1 C

[SpamAssassin Home](#) [Home](#) [News](#) [Wiki](#) [Download](#) [FAQ](#) [Docs](#) [Lists](#) [Tests](#)

Note

This is the home page for the open-source Apache SpamAssassin Project. T numerous prepackaged [versions for Windows](#), [commercial versions](#), and [front-ends](#).

ent here because you received an e-mail message which was m
n, please read [this page](#).




Handwritten notes in German, likely related to the example applications or the SpamAssassin project.

Learn from access patterns



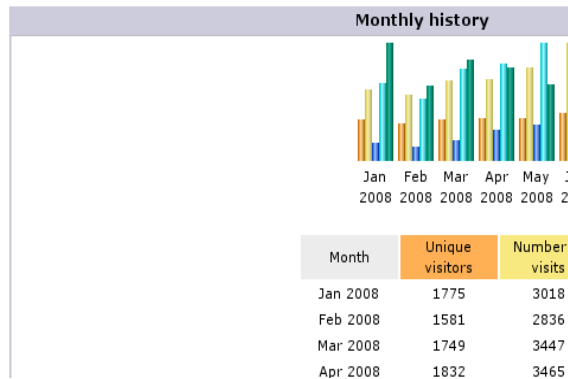
Days of week
Hours
Who:
Countries
 Full list
 Regions
 Cities
Hosts
 Full list
 Last visit
 Unresolved IP Address
Authenticated users
 Full list
 Last visit
Robots/Spiders visitors
 Full list
 Last visit
Navigation:
Visits duration
File type
Viewed
 Full list
 Entry
 Exit
Operating Systems
 Versions
 Unknown
Browsers
 Versions
 Unknown
Screen sizes
Referrers:
Origin
 Referring search engines
 Referring sites
Search

Reported period: Aug 2008 OK

 **Summary**

Reported period	Month Aug 2008	
First visit	01 Aug 2008 - 00:17	
Last visit	20 Aug 2008 - 05:06	
	Unique visitors	Number of visits
Viewed traffic *	584	913 (1.56 visits/visitor)
Not viewed traffic *		

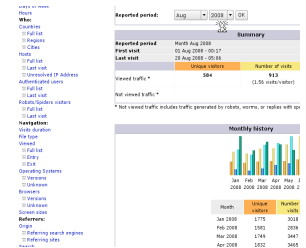
* Not viewed traffic includes traffic generated by robots, worms, or replies with sps



Learn from access patterns



- Input: Files containing server access logs.
- Some interesting tasks:
 - Identify user groups based interest in topics on site.
 - Adjust navigation to rules such as: “People who visit Debian and Mac Book pages also visit the refit pages”.
 - Show different sites to potential developers compared to visitors that are users seeking help.
- So far, a lot is done manually.



Clustering news stories



News

Top Stories

World

U.S.

Business

Elections

Sci/Tech

Entertainment

Sports

>Olympics

Health

Most Popular

☒ News Alerts

[RSS](#) | [Atom](#)
[About Feeds](#)

[Mobile News](#)

[About](#)
[Google News](#)

Olympics



[China Daily](#)

Johnson ends gold drought with Olympic win in balance beam

[Kansas City Star](#) - 51 minutes ago

US gymnast Shawn Johnson had a big smile and a gold medal at the end of Tuesday's balance beam. B EIJING | Jerry Seinfeld always says that the silver medal at the Olympics has to be the worst one.

[+Video: Olympics '08: Johnson, Liukin Take More Medals](#)
[AssociatedPress](#)

[Iowa's silver belle spins gold](#) [Minneapolis Star Tribune](#)

[Seattle Times](#) - [The Associated Press](#) - [USA Today](#) -

[Kansas City Star](#)

[all 2,969 news articles »](#)



[Canoe.ca](#)

The Making Of China's Olympic Golden Age

[CBS News](#) - 43 minutes ago

China is obsessed with Olympic gold that it is training 200000 handpicked kids in state-run sports boarding schools. (CBS) CBS News staffers file insider impressions and share their experiences throughout the day.

[Americans' necks full of Olympic medals, but precious few are gold](#)

[Dallas Morning News](#)

[Chinatown gets a lift from Beijing Olympics](#) [San Francisco Chronicle](#)

Clustering news stories



- Top Stories
- World
- U.S.
- Business
- Elections
- Sci/Tech
- Entertainment
- Sports
- >Olympics**
- Health
- Most Popular

☒ News Alerts

[RSS](#) | [Atom](#)
[About Feeds](#)

[Mobile News](#)

[About](#)
[Google News](#)

Olympics



[China Daily](#)

Johnson ends gold drought with Olympic win in balance beam

Kansas City Star - 51 minutes ago

US gymnast Shawn Johnson had a big smile and a gold medal at the end of Tuesday's balance beam. B EIJING | Jerry Seinfeld always says that the silver medal at the Olympics has to be the worst one.

[Video: Olympics '08: Johnson, Liukin Take More Medals](#)

[Iowa's silver ball spins gold](#) [Minneapolis Star Tribune](#)
[Seattle Times](#) - [The Associated Press](#) - [USA Today](#) -

[Kansas City Star](#)
[all 2,969 news articles »](#)



[Canoe.ca](#)

The Making Of China's Olympic Golden Age

CBS News - 43 minutes ago

China is obsessed with Olympic gold that it is training 200000 handpicked kids in state-run sports boarding schools. (CBS) CBS News staffers file insider impressions and share their experiences throughout the day.

[Americans' necks full of Olympic medals, but precious few are gold](#)
[Dallas Morning News](#)

[Chinatown gets a lift from Beijing Olympics](#) [San Francisco Chronicle](#)

Identify emerging
hot topics from news

Mail filtering



The Apache

The Powerful #1 C

[SpamAssassin Home](#) [Home](#) [News](#) [Wiki](#) [Download](#) [FAQ](#) [Docs](#) [Lists](#) [Tests](#)

Note

This is the home page for the open-source Apache SpamAssassin Project. T numerous prepackaged [versions for Windows, commercial versions, and front-ends](#).

If you were sent here because you received an e-mail message which was marked as spam by SpamAssassin, please read [this page](#).

Latest News

Mail filtering – Special problems



- Spam mails change over time: Adapt to filters.
 - Images that contain the text.
 - Text is modified.
- What is spam for me, might be ham for you.
- Usually users are not willing to provide labels.



The Apache
The Powerful #1 C

[SpamAssassin Home](#) [Home](#) [News](#) [Wiki](#) [Download](#) [FAQ](#) [Docs](#) [Lists](#) [Tests](#)

Note

This is the home page for the open-source Apache SpamAssassin Project. T numerous prepackaged [versions for Windows, commercial versions, and front-ends](#).

If you were sent here because you received an e-mail message which was marked as spam by SpamAssassin, please read [this page](#).

Latest News

Recommend videos



Family Filter: ON

Related: [barack obama](#) [obama berlin](#) [bush](#) [the one obama](#) [obama song](#) [p](#)

Videos

Channels

Groups

People

Deta

1 - 20 of 1478

1

2

3

4

5

6

7

...

Sort By

Most Relevant

Most Popular

Most Recent

▼ Added

All Time

Today

This Week

This Month

▼ Run Length

Time 0 - ∞ min

0 5 20 40 60 ∞

▼ Category

All

Animation

Anime

Autos & Vehicles

Search results for "obama"

Clinton/Obama: Monica --The Debate for the Black Vote (Part 3)
Monica's name comes up in PART 3 of "Clinton/Obama: The Debate for the...
★ ★ ★ ★ ★ 01:44

pro BriteThoi
added: 1 year
category: Con
views: 5895
source: Veoh.

Obama Mania Manchester Part 3
Barack Obama's initial pre-presidential election foray into New Hampsh...
★ ★ ★ ★ ★ 07:11

pro alcannist
added: 1 year
category: Ente
views: 535
source: Veoh.

Obama Mania Manchester Part 2
Barack Obama's initial pre-presidential election foray into New Hampsh...
★ ★ ★ ★ ★ 09:56

pro alcannist
added: 1 year
category: Ente
views: 359
source: Veoh.

Obama Mania Manchester Part 1
Barack Obama's initial pre-presidential election foray into New Hampsh...
★ ★ ★ ★ ★

pro alcannist
added: 1 year
category: Ente

Recommend videos



Search results for "obama"

Search: obama Search

Family Filter: ON

Related: [barack obama](#) [obama berlin](#) [bush](#) [the one obama](#) [obama song](#) [p](#)

Videos Channels Groups People Deta

1 - 20 of 1478 1 2 3 4 5 6 7 ...

Sort By
Most Relevant
Most Popular
Most Recent

▼ Added
All Time
Today
This Week
This Month

▼ Run Length
Time 0 - ∞ min
0 5 20 40 60 ∞

▼ Category
All
Animation
Anime
Autos & Vehicles

Search results for "obama"

Clinton/Obama: Monica --The Debate for the Black Vote (Part 3)
Monica's name comes up in PART 3 of "Clinton/Obama: The Debate for the..."
★★★★★

Obama Mania Manchester Part 2
Barack Obama's initial pre-presidential election foray into New Hampsh...
★★★★★ 07:11

Obama Mania Manchester Part 2
Barack Obama's initial pre-presidential election foray into New Hampsh...
★★★★★ 09:56

Obama Mania Manchester Part 1
Barack Obama's initial pre-presidential election foray into New Hampsh...
★★★★★

pro BriteThoi added: 1 year category: Con source: Veoh.

pro alcannist added: 1 year category: Ente views: 535 source: Veoh.

pro alcannist added: 1 year category: Ente views: 359 source: Veoh.

pro alcannist added: 1 year category: Ente

Provide ratings, where unavailable, infer from usage.

Recommend videos



Search results for "obama"

Search: obama Search

Family Filter: ON

Related: [barack obama](#) [obama berlin](#) [bush](#) [the one obama](#) [obama song](#) [p](#)

Videos Channels Groups People Deta

1 - 20 of 1478 1 2 3 4 5 6 7 ...

Sort By
Most Relevant
Most Popular
Most Recent

▼ Added
All Time
Today
This Week
This Month

▼ Run Length
Time 0 - ∞ min
0 5 20 40 60 ∞

▼ Category
Animation
Anime
Autos & Vehicles

Search results for "obama"

Clinton/Obama: Monica --The Debate for the Black Vote (Part 3)
Monica's name comes up in PART 3 of "Clinton/Obama: The Debate for the..."
★★★★★
pro BriteThoi added: 1 year category: Con source: Veoh.

Obama Mania Manchester Part 2
Barack Obama's initial pre-presidential election foray into New Hampsh...
★★★★★ 07:11
pro alcannist added: 1 year category: Ente views: 535 source: Veoh.

Obama Mania Manchester Part 2
Barack Obama's initial pre-presidential election foray into New Hampsh...
★★★★★ 09:56
pro alcannist added: 1 year category: Ente views: 359 source: Veoh.

Obama Mania Manchester Part 1
Barack Obama's initial pre-presidential election foray into New Hampsh...
pro alcannist added: 1 year category: Ente

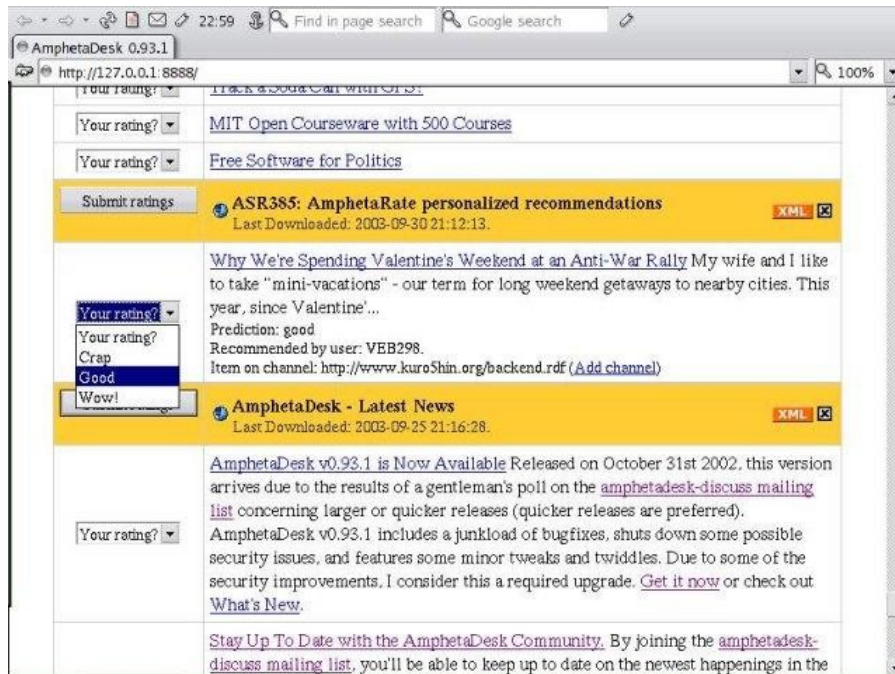
Categorize content.

Provide ratings, where unavailable, infer from usage.

Recommending RSS feeds



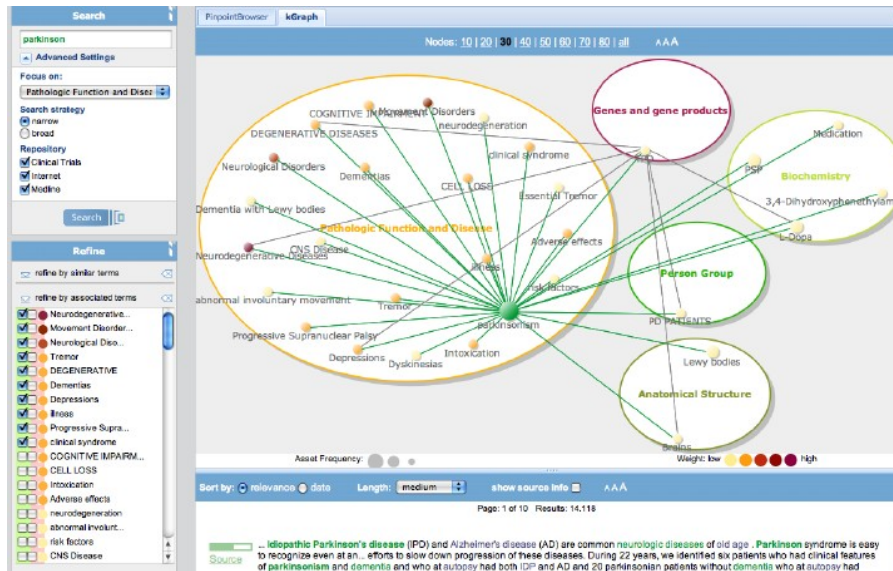
- Input:
 - List of RSS feeds.
 - Maybe ratings for feeds.
- Task:
 - Get me the latest and greatest that I like.



Recommend new papers



- Input:
 - A list of papers I like.
 - Maybe my own papers.
 - List of new publications.
- Task:
 - Give me all papers relevant for me.



Provide debugging help



Tasks	Error Log	Rose X	Progress			
Action	Symbol	File	Support	Confidence		
ADD_IN	BirdieProcessor.java	{softevo-test}\...{birdie}\BirdieProcessor.java	4	0.3333		
CHG	~usage()	{softevo-test}\...{birdie}\Birdie.java	4	0.3333		
DEL_IN	URLUtils.java	{softevo-test}\...{util}\URLUtils.java	3	0.25		
ADD_IN	URLUtils.java	{softevo-test}\...{util}\URLUtils.java	3	0.25		
CHG	~processFile(String, BirdieProcessor)	{softevo-test}\...{birdie}\Birdie.java	3	0.25		
CHG	~close()	{softevo-test}\...{birdie}\BirdieProcessor.java	3	0.25		
ADD_IN	ArrayUtils.java	{softevo-test}\...{util}\ArrayUtils.java	2	0.1666		

```
StandardSourcePathProvider.java
} else {
    // recover persisted source path
    entries = recoverRuntimePath(con
}
return entries;
}

/* (non-Javadoc)
 * @see IRuntimeClasspathEntry[] resolve()
 */
public IRuntimeClasspathEntry[] resolve()
{
    List all = new ArrayList<>(entries.length);
    for (int i = 0; i < entries.length; i++)
    {
        switch (entries[i].getType())
        {
            case IRuntimeClasspathEntry.
                // a project resolves to
                all.add(entries[i]);
                break;
            case IRuntimeClasspathEntry.
                IRuntimeClasspathEntry2
                String typeId = entry.ge
                IRuntimeClasspathEntry[]
                if (typeId.equals(Default
                // add the resolved
        }
    }
}
```

eROSE: Guiding Programmers in Eclipse:
<http://www.st.cs.uni-sb.de/softevo/erose/>

Provide debugging help



Mark pieces that are edited together.

Tasks	Error Log	Rose	Progress			
Action	Symbol	File	Support	Confidence		
ADD_IN	BirdieProcessor.java	{softevo-test}...\birdie\BirdieProcessor.java	4	0.3333		
CHG	~usage()	{softevo-test}...\birdie\Birdie.java	4	0.3333		
DEL_IN	URLUtils.java	{softevo-test}...\util\URLUtils.java	3	0.25		
ADD_IN	URLUtils.java	{softevo-test}...\util\URLUtils.java	3	0.25		
CHG	~processFile(String, BirdieProcessor)	{softevo-test}...\birdie\Birdie.java	3	0.25		
CHG	~close()	{softevo-test}...\birdie\BirdieProcessor.java	3	0.25		
ADD_IN	ArrayUtils.java	{softevo-test}...\util\ArrayUtils.java	2	0.1666		

```
StandardSourcePathProvider.java
} else {
    // recover persisted source path
    entries = recoverRuntimePath(con
}
return entries;
}

/* (non-Javadoc)
 * @see IRuntimeClasspathEntry[] resolve()
 */
public IRuntimeClasspathEntry[] resolve()
{
    List all = new ArrayList(entries.length);
    for (int i = 0; i < entries.length; i++)
    {
        switch (entries[i].getType())
        {
            case IRuntimeClasspathEntry.
                // a project resolves to
                all.add(entries[i]);
                break;
            case IRuntimeClasspathEntry.
                IRuntimeClasspathEntry2
                String typeId = entry.ge
                IRuntimeClasspathEntry[]
                if (typeId.equals(Default
                // add the resolved
        }
    }
}
```

eROSE: Guiding Programmers in Eclipse:
<http://www.st.cs.uni-sb.de/softevo/erose/>



Provide debugging help

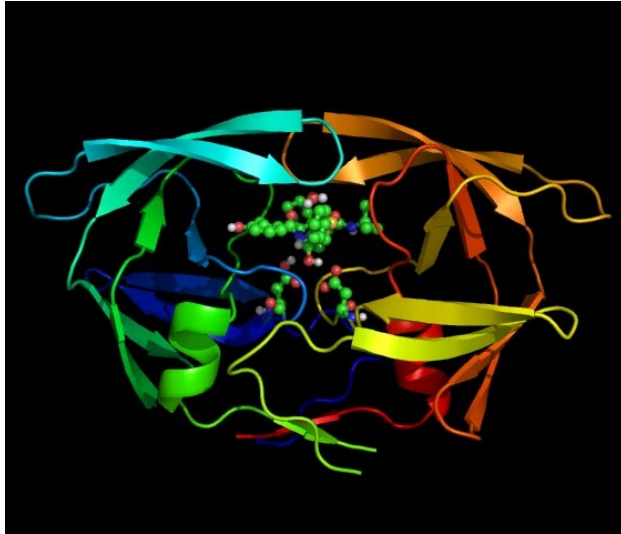
Mark pieces that are edited together.

Action	Symbol	File	Support	Confidence
ADD_IN	BirdieProcessor.java	{softevo-test}...\{birdie\BirdieProcessor.java	4	0.3333
CHG	~usage()	{softevo-test}...\{birdie\Birdie.java	4	0.3333
DEL_IN	URLUtils.java	{softevo-test}...\{util\URLUtils.java	3	0.25
ADD_IN	URLUtils.java	{softevo-test}...\{util\URLUtils.java	3	0.25
CHG	~processFile(String, BirdieProcessor)	{softevo-test}...\{birdie\Birdie.java	3	0.25
CHG	~close()	{softevo-test}...\{birdie\BirdieProcessor.java	3	0.25
ADD_IN	ArrayUtils.java	{softevo-test}...\{util\ArrayUtils.java	2	0.1666

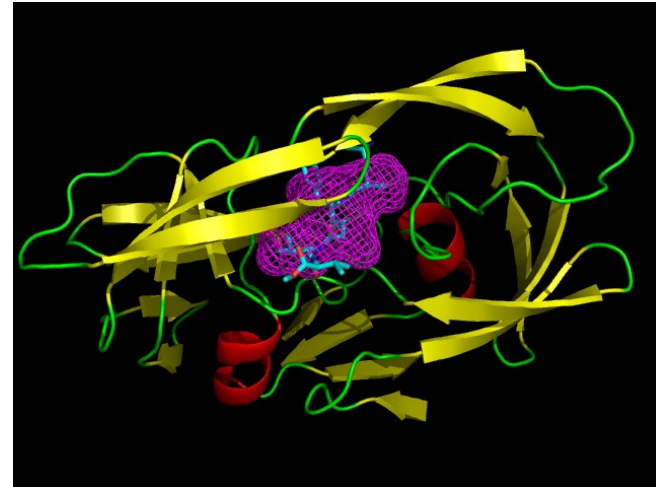
Mark pieces that are related to bugs.

eROSE: Guiding Programmers in Eclipse:
<http://www.st.cs.uni-sb.de/softevo/erose/>

Find new medications



Analyse the HIV virus to predict drug resistancies.



Make job searches easier



Suche verfeinern

Anlagentechnik

Berater

Business

[zeige alle 11 Filter](#)

Region

Bonn [120]

Troisdorf [8]

Siegburg [7]

[zeige alle 11 Regionen](#)

Tätigkeitsbereiche

Informationst... [112]

Technische Tä... [22]

Consulting & ... [18]

[zeige alle 11 Bereiche](#)

Position

Angestellter/... [145]

Gruppenleiter... [9]

Praktikant [4]

[zeige alle 8 Positionen](#)

Seite: **1** 2 3 4 5 6 7 | [nächste Seite](#)

Hier könnte Ihre Stellenanzeige stehen.

Jetzt Top-Platzierung b

IS Engineer / Business Engineer

Arbeitgeber: [Koerschgens GmbH](#) in [Bonn](#) 01.1

IS Engineer / Business Engineer it2work IS Engineer / Business Engineer Nummer: AUS90062
Beschäftigungsbeginn: 01.01.2008 Ihre Verantwortung Business Engineering und IT-Teilprojektleitungen i
Auftrag der Projektmanager, Business- und Systemowner. Requirementengineering, -Management [...] [m](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer (m/w)

Arbeitgeber: [PWB-Ruhlatec Industrieprodukte GmbH](#) in [Sankt Augustin](#) 16.1

[...] Aktuelle Stellenangebote finden Sie hier: Sales Engineer (m/w) Elektroniker (m/w) Sales Engineer (m/w)
Ihren Aufgaben gehören: Neuakquise von Kunden und die Ermittlung deren Leistungsprofils Betreuung de
bestehenden Kundenstamms Präsentation von technisch [...] [mehr Info](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer / Vertriebsingenieur (m/w)

Arbeitgeber: [personal total](#) in [Bonn](#) 10.1

Make job searches easier



Suche verfeinern

Anlagentechnik

Berater

Business

[zeige alle 11 Filter](#)

Region

Bonn [120]

Troisdorf [8]

Siegburg [7]

[zeige alle 11 Regionen](#)

Tätigkeitsbereiche

Informationst... [112]

Technische Tä... [22]

Consulting & ... [18]

[zeige alle 11 Bereiche](#)

Position

Angestellter/... [145]

Gruppenleiter... [9]

Praktikant [4]

[zeige alle 8 Positionen](#)

Seite: **1** 2 3 4 5 6 7 | [nächste Seite](#)

Hier könnte Ihre Stellenanzeige stehen. [Jetzt Top-Platzierung bei](#)

IS Engineer / Business Engineer

Arbeitgeber: [Koerschgens GmbH in Bonn](#) 01.1

IS Engineer / Business Engineer it2work IS Engineer / Business Engineer Nummer: AUS90062
Beschäftigungsbeginn: 01.01.2008 Ihre Verantwortung Business Engineering und IT-Teilprojektleitungen i
Auftrag der Projektmanager, Business- und Systemowner. Requirementengineering, -Management [...] [m](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer (m/w)

Arbeitgeber: [PWB-Ruhlatec Industrieprodukte GmbH in Sankt Augustin](#) 16.1

[...] Aktuelle Stellenangebote finden Sie hier: Sales Engineer (m/w) Elektroniker (m/w) Sales Engineer (m/w)
Ihren Aufgaben gehören: Neuakquise von Kunden und die Ermittlung deren Leistungsprofils Betreuung de
bestehenden Kundenstamms Präsentation von technisch [...] [mehr Info](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer / Vertriebsingenieur (m/w)

Arbeitgeber: [personal total in Bonn](#) 10.1

Find job postings
on the internet.

Make job searches easier



Suche verfeinern

Anlagentechnik ☐

Berater ☐

Business ☐

[zeige alle 11 Filter](#)

Region

Bonn [120] ☐

Troisdorf [8] ☐

Siegburg [7] ☐

[zeige alle 11 Regionen](#)

Tätigkeitsbereiche

Informationst... [112] ☐

Technische Tä... [22] ☐

Consulting & ... [18] ☐

[zeige alle 11 Bereiche](#)

Position

Angestellter/... [145] ☐

Gruppenleiter... [9] ☐

Praktikant [4] ☐

[zeige alle 8 Positionen](#)

Seite: **1** 2 3 4 5 6 7 | [nächste Seite](#)

Hier könnte Ihre Stellenanzeige stehen. [Jetzt Top-Platzierung](#)

IS Engineer / Business Engineer

Arbeitgeber: [Koerschgens GmbH in Bonn](#) 01.1

IS Engineer / Business Engineer it2work IS Engineer / Business Engineer Nummer: AUS90062
Beschäftigungsbeginn: 01.01.2008 Ihre Verantwortungen: Business Engineering und IT-Teilprojektleitungen i
Auftrag der Projektmanager, Business- und Systemowner, Requirementengineering, -Management [...] [m](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer (m/w)

Arbeitgeber: [PWB-Ruhlatec Industrieprodukte GmbH in Sankt Augustin](#) 16.1

[...] Aktuelle Stellenangebote finden Sie hier: Sales Engineer (m/w) Elektroniker (m/w) Sales Engineer (m/w)
Ihren Aufgaben gehören: Neukundengewinnung und die Ermittlung deren Leistungsprofils Betreuung der
bestehenden Kundenstamms Präsentation von technisch [...] [mehr Info](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer / Vertriebsingenieur (m/w)

Arbeitgeber: [personal total in Bonn](#) 10.1

Find job postings
on the internet.

Automatically extract:
Place, title, date...

Make job searches easier



Suche verfeinern

Anlagentechnik ☐

Berater ☐

Business ☐

[zeige alle 11 Filter](#)

Region

Bonn [120] ☐

Siegburg [7] ☐

[zeige alle 11 Regionen](#)

Tätigkeitsbereiche

Informationst... [112] ☐

Technische Tä... [22] ☐

Consulting & ... [18] ☐

[zeige alle 11 Bereiche](#)

Position

Angestellter/... [145] ☐

Gruppenleiter... [9] ☐

Praktikant [4] ☐

[zeige alle 8 Positionen](#)

Group by categories.

Seite: **1** 2 3 4 5 6 7 | [nächste Seite](#)

Hier könnte Ihre Stellenanzeige stehen. [Jetzt Top-Platzierung](#)

IS Engineer / Business Engineer

Arbeitgeber: [Koerschgens GmbH in Bonn](#)

IS Engineer / Business Engineer it2work IS Engineer / Business Engineer Nummer: AUS90062
Beschäftigungsbeginn: 01.01.2008 Ihre Verantwortungen: Business Engineering und IT-Teilprojektleitungen i
Auftrag der Projektmanager, Business- und Systemowner, Requirementengineering, -Management [...] [m](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

Sales Engineer (m/w)

Arbeitgeber: [PWB-Ruhlatec Industrieprodukte GmbH in Sankt Augustin](#)

[...] Aktuelle Stellenangebote finden Sie hier: Sales Engineer (m/w) Elektroniker (m/w) Sales Engineer (m/w)
Ihren Aufgaben gehören: Neukakquise von Kunden und die Ermittlung deren Leistungsprofils Betreuung de
bestehenden Kundenstamms Präsentation von technisch [...] [mehr Info](#)

[Web](#) [Treffer in Karte](#) [Ähnliche Angebote](#) [Job merken](#)

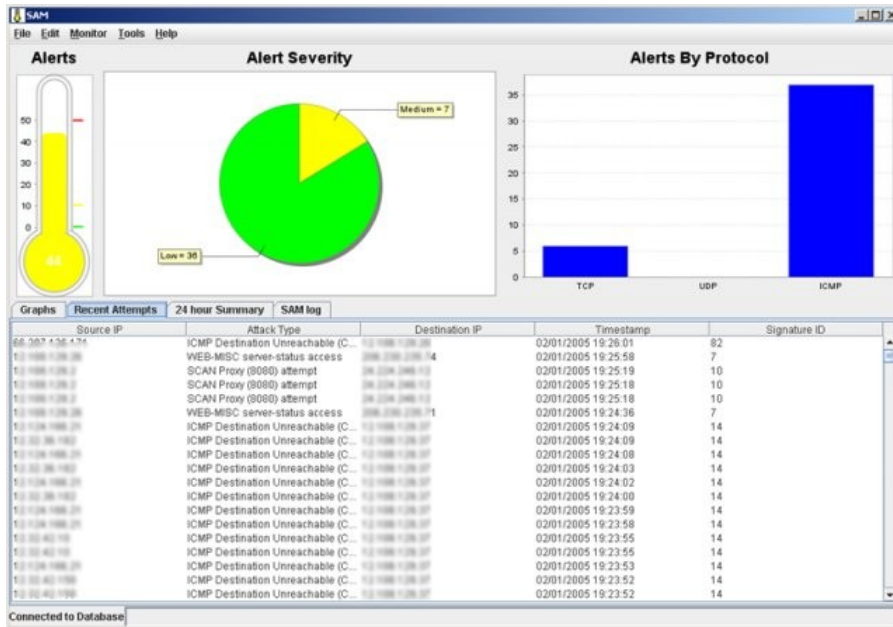
Sales Engineer / Vertriebsingenieur (m/w)

Arbeitgeber: [personal total in Bonn](#)

Find job postings
on the internet.

Automatically extract:
Place, title, date...

Identify intrusion patterns



- Input logs for:
 - Normal users.
 - Detected attacks.
- Task:
 - Given live logs, alert me, if someone is attacking.

Learn search rankings



- Input:
 - User votings.
 - Clicks on search results.
 - Query refinement logs.
- Task:
 - Create a perfect ranking.

The screenshot shows the sproose search engine interface. At the top, there's a navigation bar with 'Web', 'Video', 'Tags', and 'Us' buttons. Below this is a search bar with 'linux' entered and a 'Search' button. A radio button selection shows 'Pages in English' is selected. Below the search bar, it says '1 - 10 of about 2741661 results found'. The results are displayed in a list format. The first result is 'Ubuntu Dedicated Servers' with a description and a link to 'http://www.ServerPronto.com/ubuntu'. The second result is 'Linux Device Software' with a description and a link to 'http://www.Handango.com'. The third result is 'Linux.com' with a description and a link to 'http://www.linux.com/'. The fourth result is 'Linux - Wikipedia, the free encyclopedia' with a description and a link to 'http://en.wikipedia.org/wiki/Linux'. The fifth result is 'Linux Kernel Archives' with a description and a link to 'http://www.kernel.org/'. Each result has a 'votes' icon (a yellow circle with a number) and a 'I like it!' link. The 'votes' icon for 'Linux.com' shows '3 votes' and '4 votes total'. The 'votes' icon for 'Linux - Wikipedia' shows '1 vote' and '2 votes total'. The 'votes' icon for 'Linux Kernel Archives' shows '1 vote' and '6 votes total'. Each result also has a 'Last Vote' section with a user profile picture, username, and number of comments.

sproose

Web Video Tags Us

Web Search: linux Search Answer Me

☐ The Web ☒ Pages in English

1 - 10 of about 2741661 results found |

Ubuntu Dedicated Servers
Dedicated Servers Give You Control. Load the Newest Ubuntu. Order Now.
[Sponsored by http://www.ServerPronto.com/ubuntu](http://www.ServerPronto.com/ubuntu)

Linux Device Software
Customize your Linux device today! Visit Handango to learn how.
[Sponsored by http://www.Handango.com](http://www.Handango.com)

Linux.com
Enterprise Linux resource with news, software, documentation, and information
<http://www.linux.com/> [more details]
Last Vote: voteman | 0 comments

Linux - Wikipedia, the free encyclopedia
Linux is one of the most prominent examples of free software and open source creator of the Linux kernel. See also: History of Linux ...
<http://en.wikipedia.org/wiki/Linux> [more details]
Last Vote: romanlagunov | 0 comments

Linux Kernel Archives
The primary site for the Linux kernel source.
<http://www.kernel.org/> [more details]
Last Vote: mortonfox | 0 comments

Aggregate information



- Input: e.g. data found in
 - Social networking sites.
 - Regular search engines.
- Task: find information
 - About some person.
 - About a company.

First Name or Initial: isabel Middle Name or Initial: Last Name: drost Keywords:

Social Networks Scientific Publications

LinkedIn IMDB facebook myspace friendster WIKIPEDIA

2 Results 0 Results 0 Results 0 Results 0 Results 0 Results

Software Engineer at Neofonie GmbH Germany

It 'zat? ♥ ♡ ✖

Dive Deeper USSEARCH

Refine your Results

Where 'zat? ---

'zat ♀♂? ---

'zat old? ---

Refine by conce

0c0885b4 apache apachecon dblp db ind

handen van wil hello jeff eastman last edited links

henze reconfigurable architectures spam

scheffer wrote

Isabel Drost

<http://www.naymz.com/search/isabel/drost/220447>

It 'zat? ♥ ♡ ✖

froscon2008: Isabel Drost

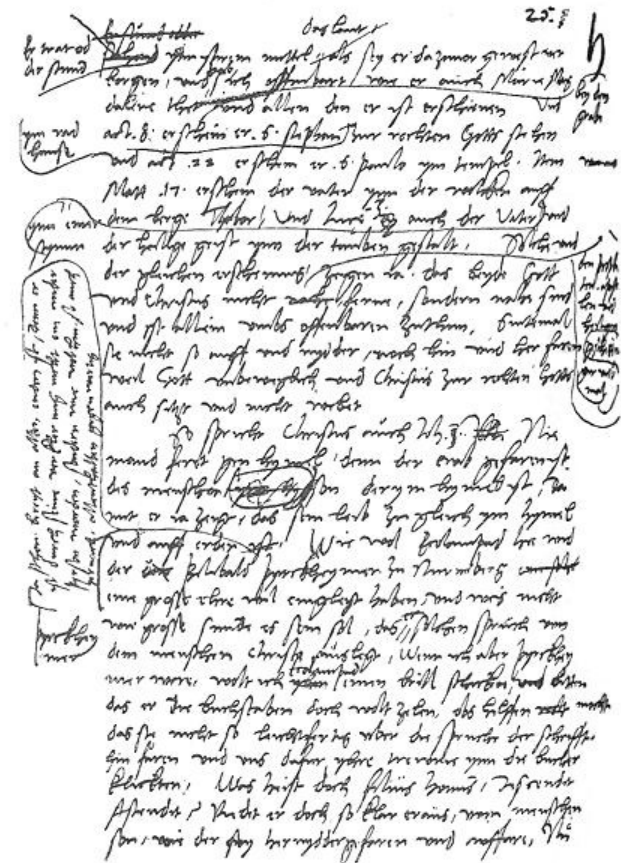
<http://programm.froscon.org/2008/speakers/201.en.html>

It 'zat? ♥ ♡ ✖

Convert handwritten text



- **Input:**
 - A handwritten text.
- **Task:**
 - Convert to (machine-) readable form.



Conclusions



- We are at the beginning.
 - High demand for scalable machine learning.
 - We need You –
 - Your enthusiasm.
 - Your mathematical knowledge.
 - Your proficiency in or will to learn Hadoop.
 - Your interest in understanding Your data.
 - mahout-dev@apache.org mahout-user@apache.org
-
-

Some advertising



Berlin - 8th of September at 5p.m.

newthinking store Berlin
Tucholskyst. 48

Hadoop User/Developer Meeting Germany

