Apache Mahout Making data analysis easy





Isabel Drost

Nighttime:

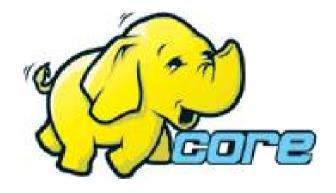
Co-Founder, committer Apache Mahout. Organiser of Berlin Hadoop Get Together.

Daytime:

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

- "Mastering Data-Intensive Collaboration and Decision Making"
- EU funded research project
 - Number of partners: 8
 - Coordinator: Research Academic Computer Technology Institute (CTI), Greece











Machine learning background?



Agenda

• Data Mining/ Machine Learning?

• Why is scaling hard?

• Going beyond simple statistics.

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.

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

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

Weight Weight – Apparent immersed weight



Nog

Wille.

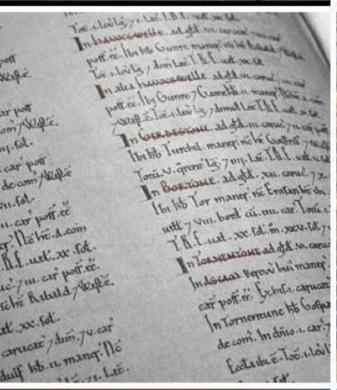
Doy u.

1.61.

A HIN

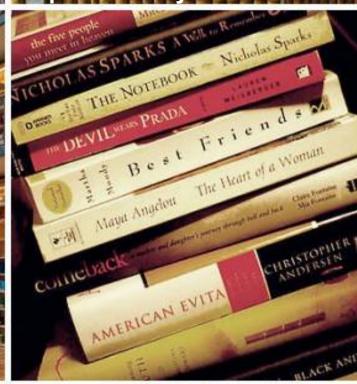




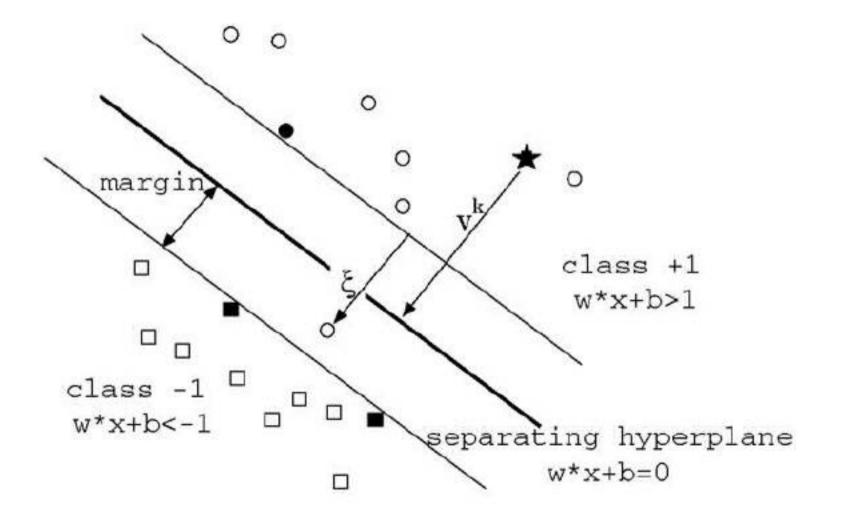




June 25, 2008 by chase-me A STREET http://www.flickr.com/photos/sasy/2609508999



An SVM's model of nature



The challenge

Mission

Provide scalable data mining algorithms.



which the manufact and she had a set of the land of the

an interfet Matury of the Interest and is incremelarly Problematic Fature

Acres 14 COLUMN TWO IS NOT

The PPOP, TPDA: and FPAR. As the form from and defines and and an and a strain of the property And and a standard of the second seco after manufa for Anothers. May the showed performance on annual or complemented and another provided processor. Every and an annual configuration processor. Every and an annual configuration processor. making the second term party for state, are word from a formal for provider the same as the system and out of an even spectrum of an even have been been der solladvance for Discontinuous, without advantage a complete testing state of a second provide second expensions and studies could be made using the same process. and successive of the denset, with the the an any post time to make the time party of an example of

property days part time area for the to death in a swell a more disking of the press and descent of the second particle press of These ordered and particulations news or except data matters in called alonger a Law, manual along a condent Manager, it prices are in the attragement concerned that it would anomales of the board on postson (doub)

http://www.flickr.com/photos/honou/2936937247/

HowTo: From data to information.

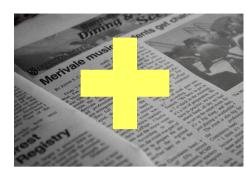
January 3, 2006 by Matt Callow http://www.flickr.com/photos/blackcustard/81680010

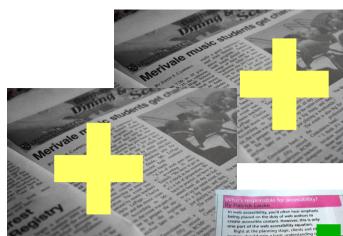
COMMUNITY NEWS

Finishing touches still to come

A glimpse of today, yesterday

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



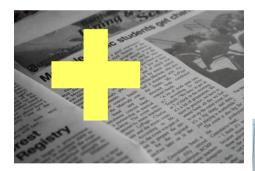






<text><text><text><text> heckpoints and current best







In web accessibility, you'll often hear emp being placed on the duty of web authors create accessible content. However, this is Dee part of the web accessibility equation Right at the planning shown

ect, and how to their project, and how to web authors/developers excellent starting point i Specification (PAS) 78 Gr

in Commissioning Ac

why it's important f

In Commissioning Access Instruct.com/utag6.endorsed by the on-Rights Commission (IRC). Web authors and developers need-comes to developing their content. The Content Accessibility Guidelines (WC WWW.W3.0rg/WA/Intro/kwag provide a solid foundation for development and evaluation. However, authors still need an actual worderstanding of the rationale underprinning orderstanding of the rationale underprinning

understanding of the rationale underpinning





Rund 7500 Demonstranten nahmen an dem Prot "Freiheit statt Angat – Stoppt den Überwachungs

experteer^{de}



experteerde 🛄

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

reachungeetaat

Development and the series of the series of the series and the series of understanding of the rationale underpinning heckpoints and current best

ticular needs.









The HDFS filesystem is not restricted to MapReduce jobs. It can be used for other applications, many of which are under way at Apache. The list includes the HBase database, the Apache Mahout machine learning system, and matrix operations.

Who's response By Patrick Lauke

Merivale mu By JUSTIN S. CAMPBELL Error of the State classi

ZEIT DATENSCHUTZ

Partnersuche

DATU!

E. QUELL

ROMM

* EMPFE

SCHLA

NEU IM R I. DATEN

im Neta

2. FUTUR URHEB

gut" . SPAM

den Hå 5. IPHON

NEU AUF

I. NACHR

Datens

Buzz

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIL SPORT

Games Internet Datenschutz Mobil

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.



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.

http://www.flickr.com/photos/topsy/204929063/ 2. CDU Un 3. GEDEN Resser

1 BUNDESLIGA Der BVB 5. DAAD Das Ende der Åra Bode





The National Aquatics Center lies on th

& Harbor , among the most imaginative architecfile projects as bullish

the great train halls Like Tempelhof, nal boasts a swee evokes the glame closing a surp

Europe, Both are part o

lew China

although at times ter

fying in their aggress

scale, they also ref

the country's effor

give shape to an en

ing national identit Foster's airport

nal, the world's

is the purest exp

of China's emb

the Modernist c

swooping form

has been con

dragon. Yet

cedent

Airport monument

conceived

Speer in 1 gateway

suggests angs placed si

GE.

Tim Griffith/PTW Architects ent ceremonial axis.

bile society that exte Grand Central Term described nation's budding wildings are

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





By Characterization of accounted by Branch accessibility parti data bare engineer create accessibility control and bare and bare and of the web accessibility control and their project, and and bare content and on what accessibility is, why it's important for web authors/developers to the buildity Analabe in Commissionity Accessibility by the buildity and their project, and how to choose up by An accessibility is, why it's important to web authors/developers to the buildity Analabe in Commissionity Accessible Webliet, and Rights Commission (RIC). Web authors and developers need to be comes to developing their context when the web authors and developers need to be comes to developing their context when a count Accessibility Guideline (WcKA) at Web authors still need an actual proderstanting of the rationale undergraining understanting of the rationale undergraining accession when the accession in the still need an actual proderstanting of the rationale undergraining and accession of the still need an actual proderstanting of the rationale undergraining accession and still need an actual proderstanting of the rationale undergraining accession accession actual accession accession



understanding of the rationale underpinning heckpoints and current best







dernet agents are a Tausende demonstrieren für Bürgerrechte im Netz particular Für einen besseren Arbeitnehmerdatenschutz die Gesundheitskarte: 130 Organisationen ha Demonstration aufgerufen. Sie fürchten den and gegen nen hatten zur

themselve

and how

understanding of the rationale underpinning



ton straitet über k CTAG S/11 Obarna

DATUM 11.9.2010 - 17.20 Uhr In QUELLE 2017 ONLINE, AFP, das IN KOMMENTARE 4 EMPEREN E-Mail verschicken | F Burr

Rund 7500 Demonstranten nahmen an dem Prot "Freiheit statt Angst – Stoppt den Überwachungs



trieren für Bürgemecht

 DATUM 11.9.2010 - 17.20 Unv
 DATUM 11.9.2010 - 17.20 Unv
 DAUELLE 2017 ONLINE, APP, doa
 NOMMENTARE 4
 MORPERLEN E-Mail verschicken | Facebook, Twitter,
 Danse Tausende demonstrieren für Bürgerrechte im Netz Für einen besseren Arbeitnehmerdatenschutz und gegen DRUCKEN Dru sationen hatten zur onstration aufgerufen. Sie fürchten der NEU IM RESSORT ZDME Kuner darf ORF-Portal Futurezone kaufen IRRECHTE "Wir haben als Kind gelennt, Teilen ia" NT PAM AUF FACEBOOK "Ein iPad umsonst, ich halte es en" LIPAD APPS Mehr Freiheit im App-Store

Patremuche Immobilien dutometit Joka Beiseangebo



ZEIT/MONLINE DATENSCHUTZ

Internet Datenachutz Mobil Garnes

die Gesundheitskarte: 130 C

Überwachungsstaa

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT NULTUR WISSEN NOTTAL STUDIUM KARRIERE LEBENDART REISEN AUTO

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





ZEIT CONLINE DATENSCHUTZ

CRUZENSON UTRAMERI ÖRZÜNCI 2008 2008000000

Internet Date

Taucondo domonetrioron für

Bürgerrechte im Netz Für einen besseren Arbeitnehmerdatenschutz und geg die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.

ORF-Portal Futurezone kaufen aben als Kind gelennt, Teilen is PAN ALE EXCEPTOR "En End unequal ich halte en i IF & IPAD APPS Mere Freiheit im Ann Store



(in Datassuche Immobilien Autometici Joka Beineurgeboie

DATUM 11.9.2010 - 17:20 Uhv
1 - QUELLE ZEIT ONLINE, AFP, dos
q KOMMENTARE 4
 EMPPERLEN E-Mail verschicken | F

RTINEL ORUCKEN Dru

stat

Rund 7500 Demonstranten nahmen an dem Prote "Freiheit statt Angst – Stoppt den Überwachungsv

reiheit

0 experteerde

B2-Particle Latation of the conservation of the part of the work according to the day of web automs the set of Content Accessibility Guidelines (WCAC) at WWWW3.012/WA/Intro/Wcag provide a solid foundation for development and evaluation. However, authors still need an actual understanding of the rationale underplinning development and current best

Description of the end of the end

understanding of the rationale underpinning heckpoints and current best







experteerde 🧟

From data to information.

Collect data and define your learning problem.

• Data preparation.

• Training a prediction model.

• Checking the performance of your model.

ZEIT CONLINE DATENSCHUTZ

Partnersuche Immobilien Automarkt Jobs Reiseangebote

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO

Internet Datenschutz Mobil Games

Anmelden | Registrieren

Suchen

DATENSCHUTZ

SPORT

Tausende demonstrieren für **Bürgerrechte im Netz**

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



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.

netsenten vendten eich unter anderem zoren die Vellezsählung Dia Da

DATUM 11.9.2010 - 17:20 Uhr

- I- QUELLE ZEIT ONLINE, AFP, dpa
- C KOMMENTARE 4
- * EMPFEHLEN E-Mail verschicken | Facebook, Twitter, Buzz .
- ARTIKEL DRUCKEN Druckversion | PDF SCHLAGWORTE Datenschutz | Demonstration |
- Datensicherheit | Medienpolitik

NEU IM RESSORT

- I. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte im Netz 2. FUTUREZONE Kurier darf ORF-Portal Futurezone kaufen
- 3. URHEBERRECHTE "Wir haben als Kind gelernt, Teilen ist
- out" 1. 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

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





ZEIT CONLINE | DATENSCHUTZ

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO

Internet Datenschutz Mobil Games

DATENSCHUTZ

SPORT

Tausende demonstrieren für Bürgerrechte im Netz

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



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

Die Domonstranten wandten eich unter anderem zeren die Vellersählung

Anmelden | Registrieren

Partnersuche Immobilien Automarkt Jobs Reiseangebote

- KOMMENTARE 4
 EMPFEHLEN E-Mail verschicken | Facebook, Twitter,
 Buzz
- ARTIKEL DRUCKEN Druckversion | PDF
 SCHLAGWORTE Datenschutz | Demonstration |
 Datensicherheit | Medienpolitik

NEU IM RESSORT I. 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" 5. SPAM AUF FACEBOOK "Ein iPad umsonst, ich halte es in den Händen" 5. IPHORE & IPAD APPS Mehr Freiheit im Apo-Store

NEU AUF ZEIT ONLINE I. NACHRUF Die Freie - Bärbel Bohley ist tot 2. CDU Union streitet über konservatives Profil

 GEDENKTAG 9/11 Obama warnt vor religiösen Ressentiments
 BUNDESLIGA Der BVB und die Freiburger siegen
 S. DAAD Das Ende der Ära Bode

ANZEIGE

• Remove noise.

ZEIT CONLINE | DATENSCHUTZ

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO

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.



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

Die Demonstranten wandten eich unter anderem zonen die Vellezeihlung

Anmelden | Registrieren

Partnersuche Immobilien Automarkt Jobs Reiseangebote

- * EMPFEHLEN E-Mail verschicken | Facebook, Twitter, Buzz ...
- ARTIKEL DRUCKEN Druckversion | PDF
 SCHLAGWORTE Datenschutz | Demonstration |
 Datensicherheit | Medienpolitik

NEU IM RESSORT I. 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" **§ 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

5. DAAD Das Ende der Ära Bode

NACHRUF Die Freie - Bärbel Bohley ist tot
 COU Union streifet über konservatives Profil
 GEDENKTAG 9/11 Obama wamt vor religiösen
 Ressentiments
 BUNDESLIGA Der BVB und die Freiburger siegen

avzeige

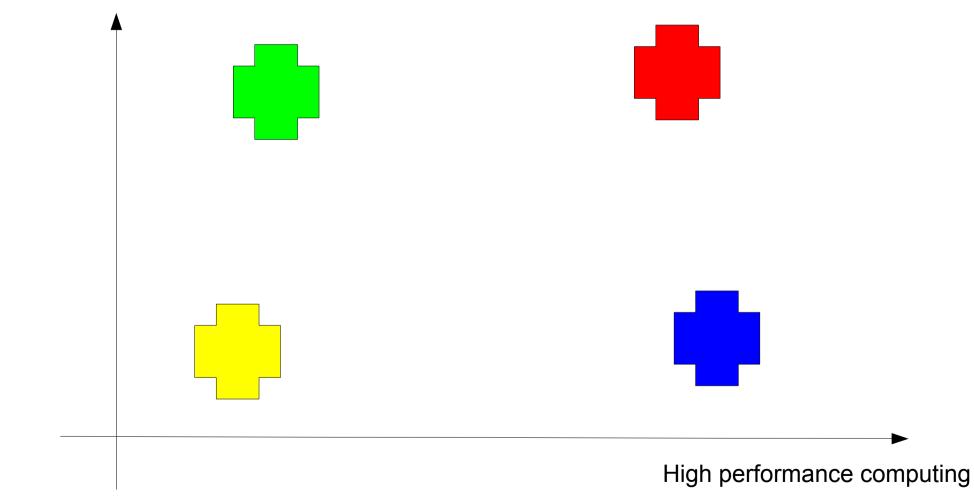
• Remove noise.

• Convert text to vectors.

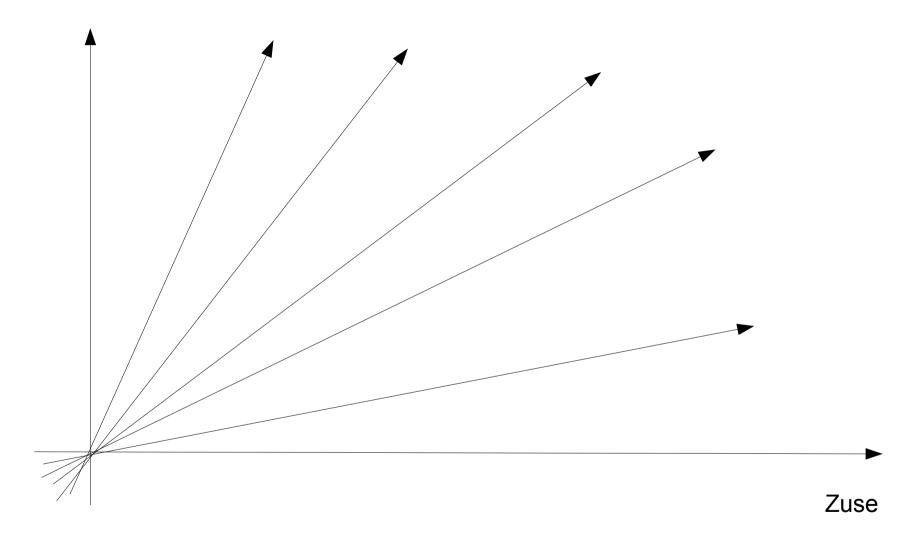
From texts to vectors

If we looked at two words only:









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.

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

- Problem:
 - Long texts get larger values.

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.

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

- Problem:
 - Additional domain knowledge ignored.

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.

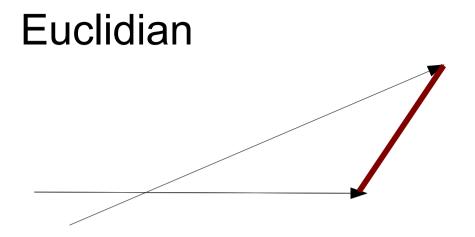
From data to information.

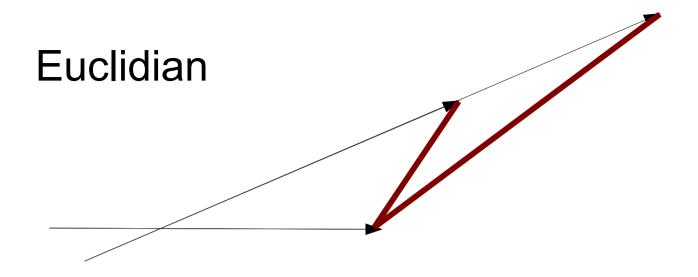
Collect data and define your learning problem.

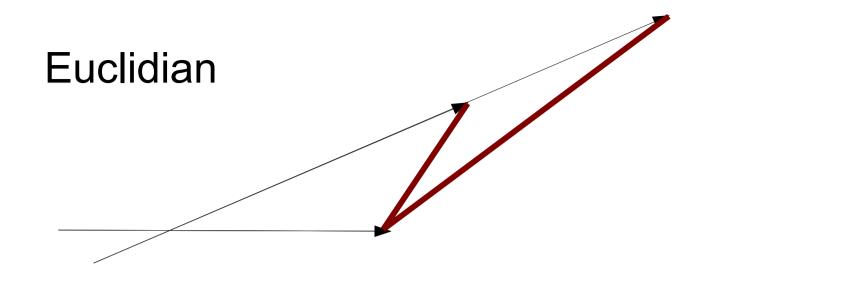
• Training a prediction model.

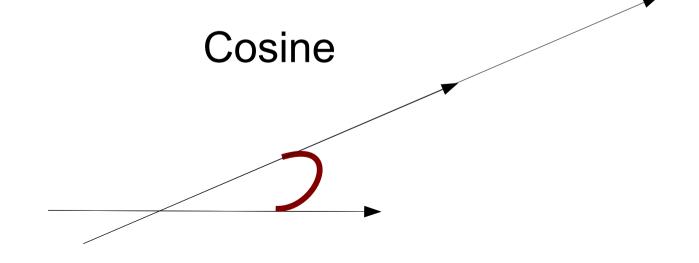
• Checking the performance of your model.

Step 2: Similarity

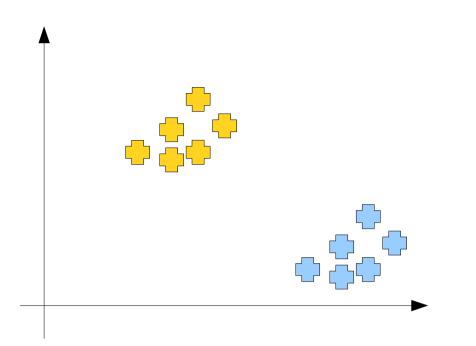


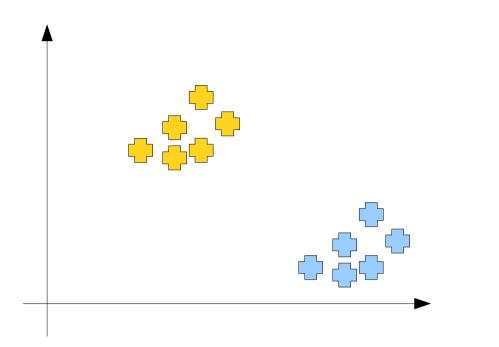


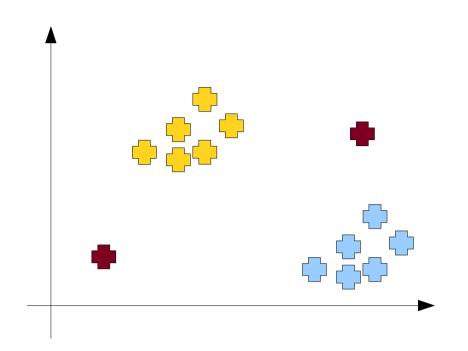


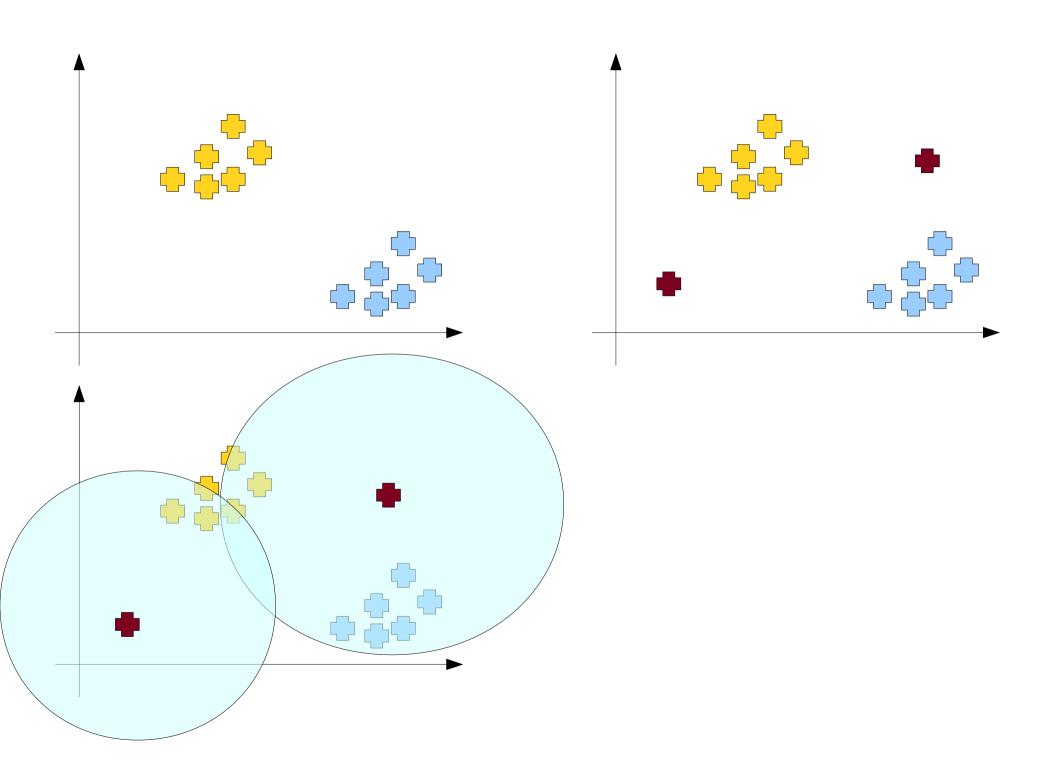


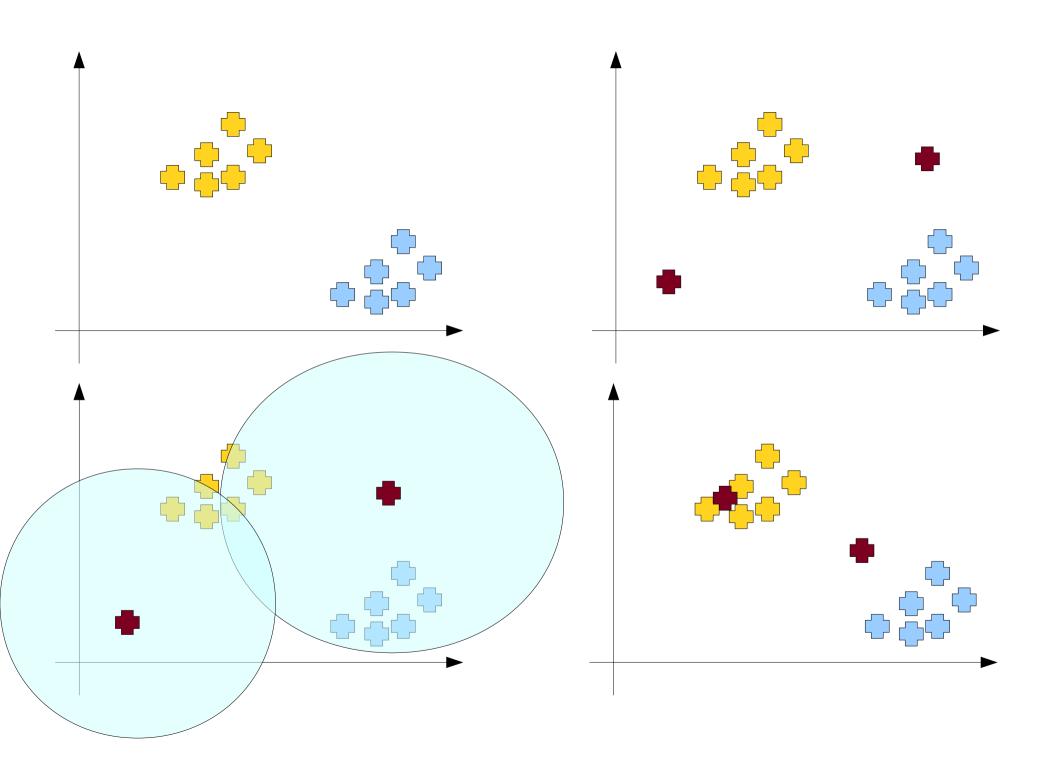
Step 3: Clustering

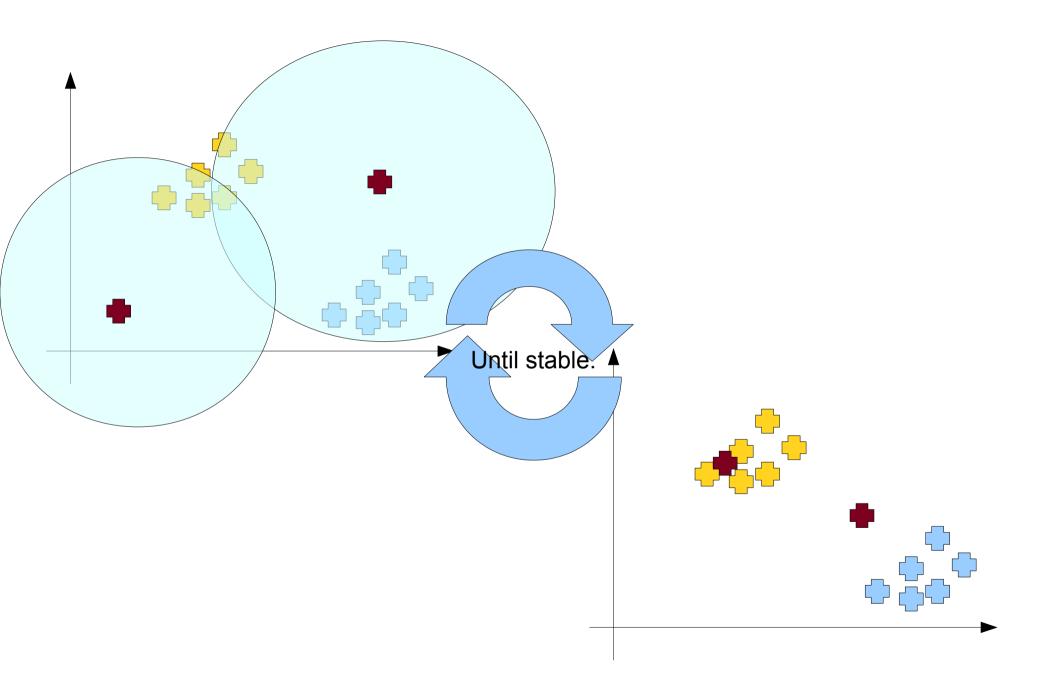












Reality

• Seed selection.

• Choice of initial k.

• Continuous updates.

• Regular addition of clusters.

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

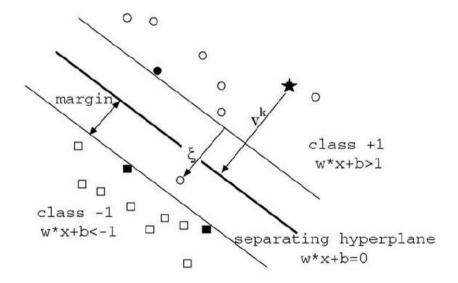
• Given a dataset of texts, identify main topics.

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.



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

SPAM

SPAM

SPAN

SPAM

SPAM

SPAN

SPAN SPAN

SPAM

Hormel

<u>SPAI</u>

SPAN

SPA

SPA

SPAM

SPAM

SPA

Leving .

Horme

SPAM

SPAM

SPAR

Google images oakland

SafeSearch: Moderate V

Related searches: oakland raiders

Images - Hide options

> 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

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



Search images

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



Advanced Image Search

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



Result

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

SafeSearch: Moderate V

Related searches: oakland raiders

Images > Face - Hide options

> Any size

Medium Large

Icon Larger than ...

Exactly

Any t Face Photo

Clip an Line drawing

Reset options

> Any color Full color Black and white Specific color



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



Search images

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



Advanced Image Search

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



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

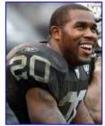


oakland@coe.ufl.edu 379 x 471 - 110k - jpg coe.ufl.edu



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

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



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



Detroit Tigers v 443 x 594 - 74k

zimbio.com



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

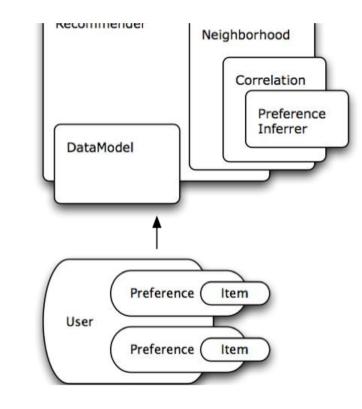
Distributed by Tubemogul. The 720 x 480 - 18k

channels.com



Recommendation mining.

• Collaborative filtering.

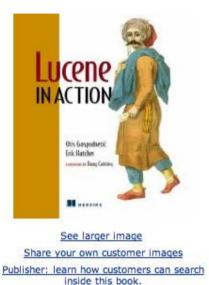


Show most relevant ads

clearasil" erwandte Suchbegriffe: <mark>clearasil pickelstift, world of warcraft.</mark>	
este Ergebniss	e Zurück Seil
Clean	Lieferung bis Samstag, 22. März: Bestellen Sie innerhalb der nächsten 23 Stunden Kostenlose Lieferung möglich: ************************************
Er	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
I.	World of WarCraft: Wrath of the Lich King (Add-on) von Vivendi Unive Vista / XP) <u>Neu kaufen</u> : EUR 39,99

Vorbestellbar Kostenlose Lieferung möglich. Games: Alle Artikel ansehen

Show most relevant ads



~ <u>Otis Gospodnetic</u> (Author), <u>Erik Hatcher</u> (Author)

List Price: \$44.95

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

In Stock.

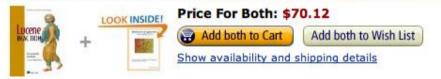
Ships from and sold by Amazon.com. Gift-wrap available.

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



Frequently Bought Together

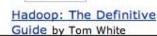
Customers buy this book with Building Search Applications: Lucene, LingPipe, and Gate by Manu Konchady



Customers Who Bought This Item Also Bought









Hibernate Search in Action by Emmanuel







Collective Intelligence in Action by Satnam Alag

Recommending places

http://www.flickr.com/photos/sebastian_bergmann/1244514498

http://www.flickr.com/photos/jfclere/4061801735





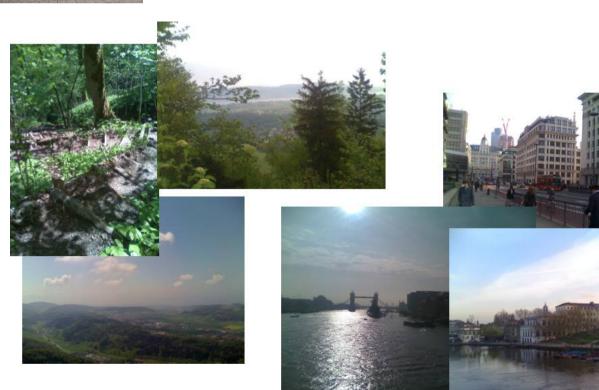








Thanks to Falko Menge for the pictures of Brussels.



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 cooccurring items.

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



rypto/3201254932/sizes/l/

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





• AWS	• Products	• Developers	··· Community	Y Support	🗸 Account	
Products & Services	 Amazon 	Elastic Compu	te Cloud (Amazo	on EC2)		
Amazon EC2 Details		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				
EC2 Overview	· · · · · · · · · · · · · · · · · · ·	ier for developers.	-	446 (1977)). 		
FAQs	Amazon EC2's	simple web service interfa	ice allows you to obtain and			
Amazon EC2 SLA	configure capa	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				
EC2 Instance Types						

Amazon Elastic MapReduce

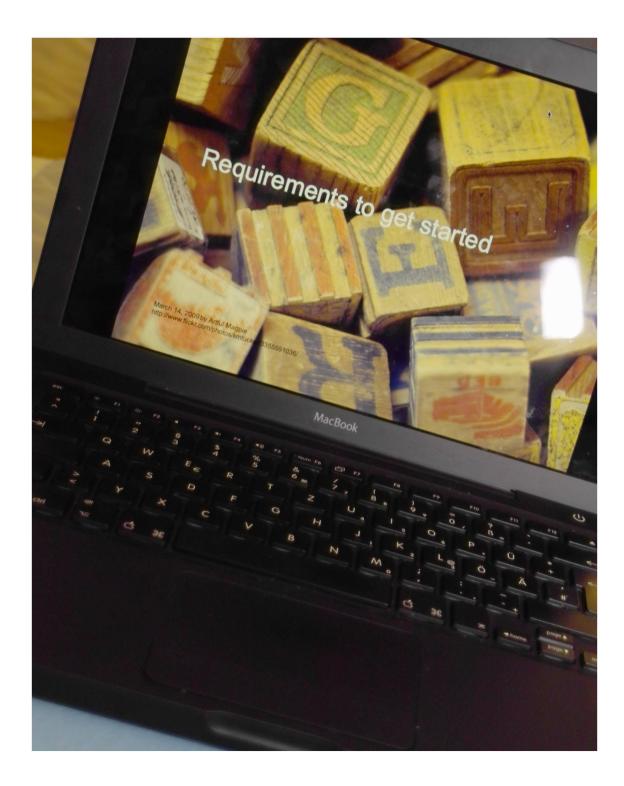
Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon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





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





http://www.flickr.com/photos/whassupbud55/4581192168









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@mahout.apache.org *-dev@mahout.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.



AND OPEN SOURCE SOFTWARE DEVELOPERS' EUROPEAN MEETING

FREE AND OFEN SUURCE SOFTWARE DEVELOPERS' EUKOPEAN MEETING

FOSDEM is the biggest free and non-commercial event organized by and for the community. Its goal is to provide Free and Open Source developers a place to meet. No registration necessary.

Search

Front page About FOSDEM Sponsors Contact

Home

Accepted devrooms

We were very pleased with the devroom proposals we received, as many were collaborative proposals with ample opportunity for cross-pollination between projects. As a side effect, this did make the decision process even harder.

Data Analytics

Open Source tools for Data Analytics. This includes Machine Learning, Text Mining and Natural Language Processing, Computer Vision, Recommender Engines, Visualisation, Data Mining, Large Scale and distributed data processing, etc.

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

ho

10

Herzlich Willkommen!

l love lelvetik a

Berlin Buzzwords 2011

Search/ Store/ Scale

May/ June 2011

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

*-user@mahout.apache.org *-dev@mahout.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.