



Open Source development for students.

Why should I work on free software?



Isabel Drost

Nighttime:

Co-Founder Apache Mahout.
Organizer of Berlin Hadoop Get Together.
Member ComDev PMC.

Daytime:

Software developer

Hello...

HPI students.

Agenda

- The Apache Software Foundation.
- Apache Mahout.
- Reasons and ways to get started.
- Invitation.



What?

Apache Software Foundation



Community over code.



Meritocracy.



Open communication.

NOT:

Github, Google Code, sourceforge.

How?

Behind the scenes.

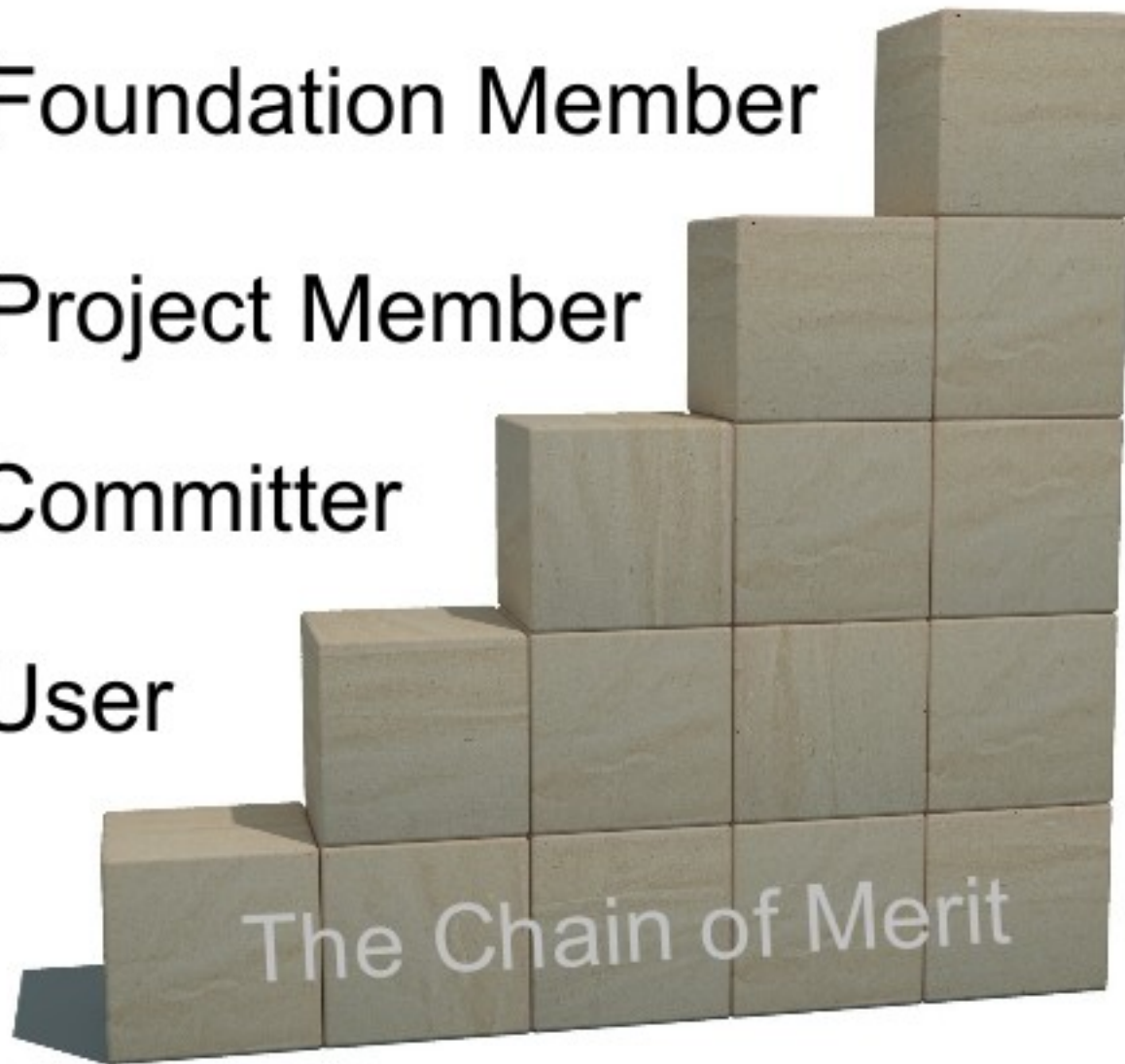
Foundation Member

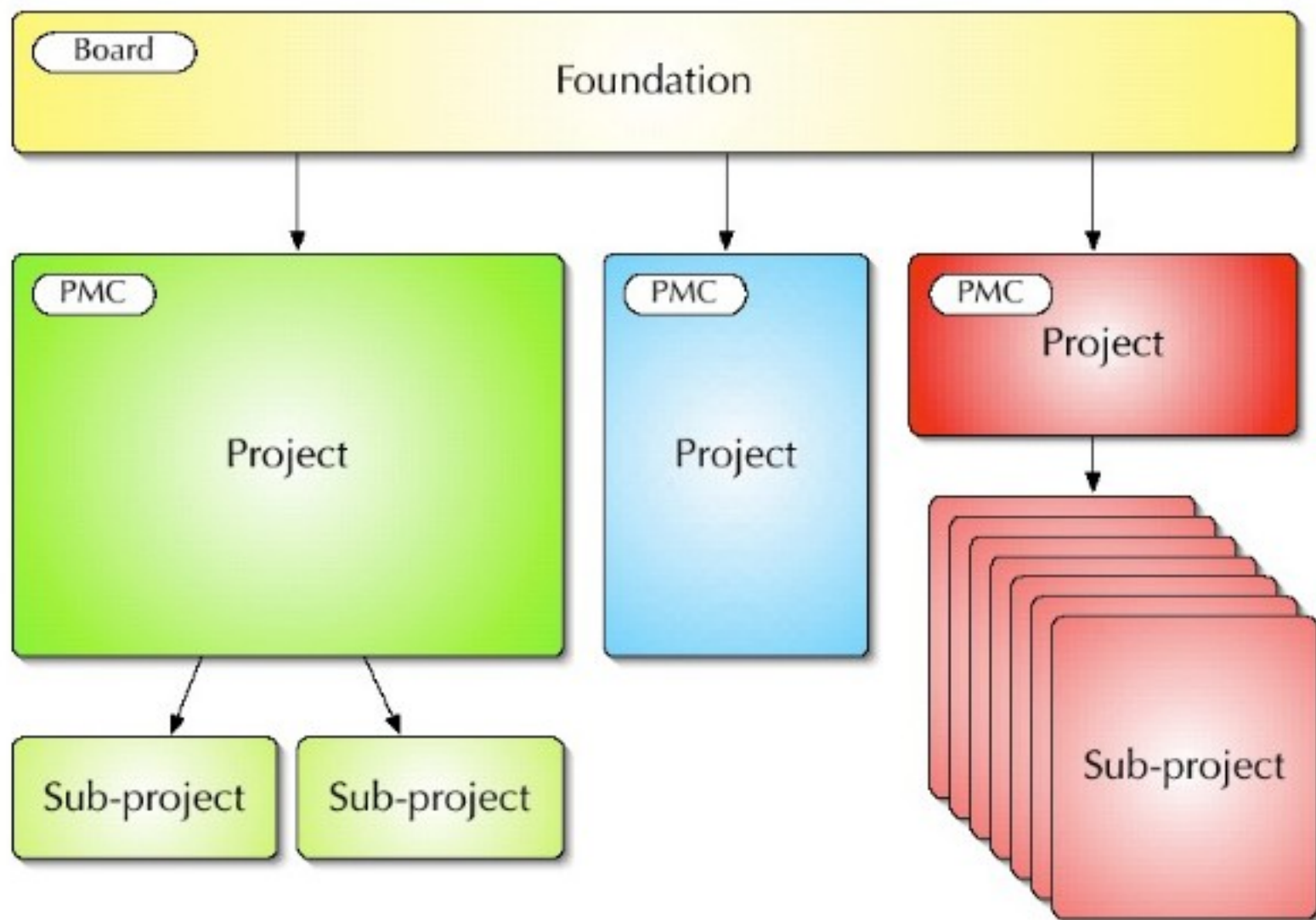
Project Member

Committer

User

The Chain of Merit





How are decisions made?

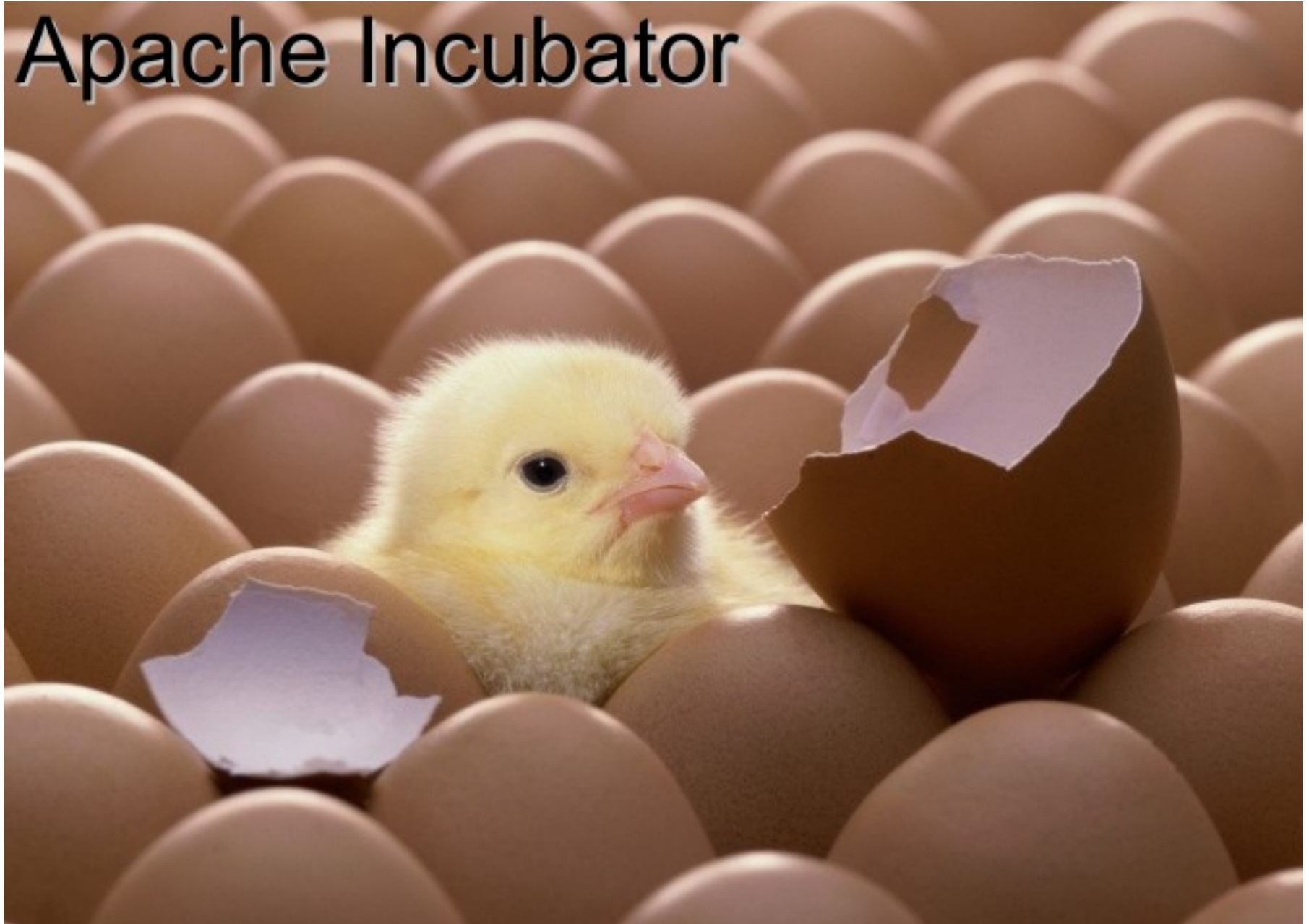
+1



-1

0

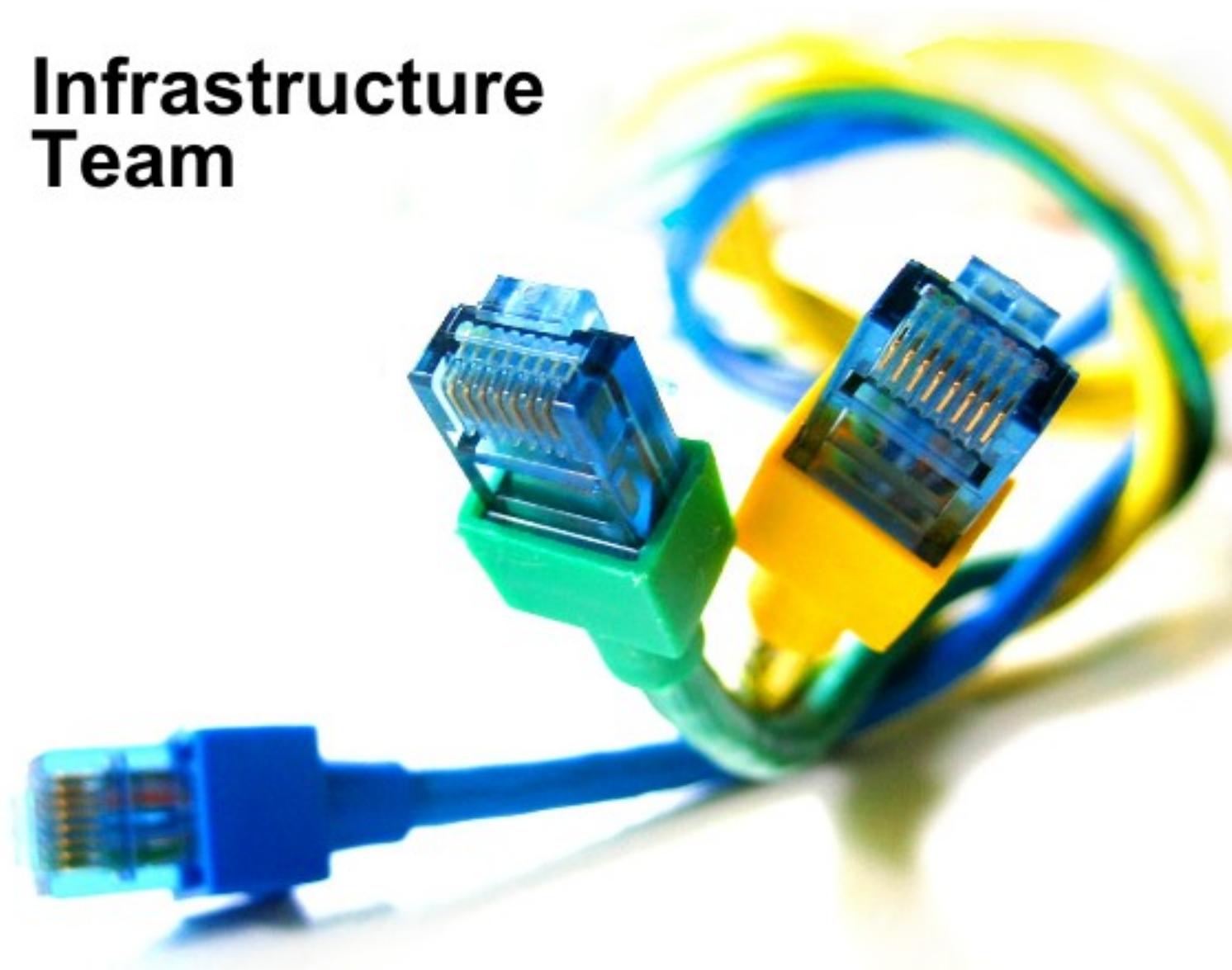
Apache Incubator



R.I.P

Apache Attic

Infrastructure Team



Apache Labs



monitoring.apache.org



© by United Feature Syndicate, Inc.

Security Response Team

security@apache.org



Public Relations Committee

press@apache.org



Fund Raising

fundraising@apache.org

Sponsorship Program

sponsor.apache.org

Platinum sponsors are:
Google, Yahoo, Microsoft



Conference Committee

feedback@apachecon.com

- ApacheBarCampOxford 2009, Oxford, UK, 4-5 Apr.
- ApacheBarCamp Asia 2009, (to be announced)
- ApacheCon US 2009, Oakland, CA, 2-6 Nov.
- ApacheCon Europe 2010, spring (to be announced)
- ApacheCon US 2010, Atlanta, GA, 1-5 Nov.
- ApacheCon Europe 2011, spring (to be announced)
- ApacheCon North America 2011, Vancouver, Canada, 7-11 Nov.
- **announce-subscribe@apachecon.com**

Travel Assistance Committee



ASF Legal Team



legal-discuss@apache.org



Java
Community
Process
MEMBER



Community development

GsoC

Mentoring

University relations

1956
committers

1100
postings/day

800
mailing lists


83,000
unique
subscribers

310
members

760,000
SVN revisions

~70
GByte in SVN

64
dev projects


33
incubator
projects

300
revisions/day

260,000,000
requests/month

1
paid sysadmin

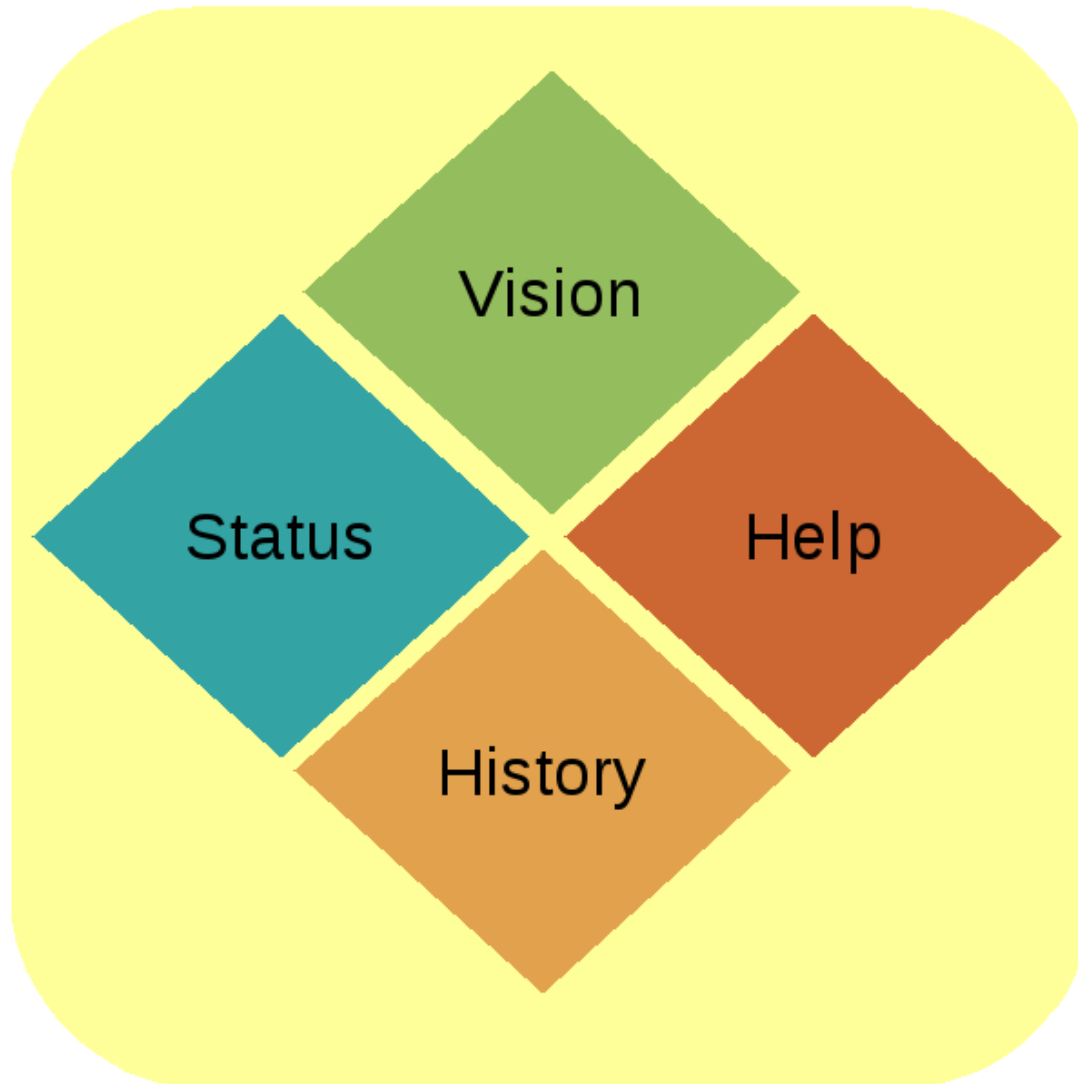
27
servers

5
data centers

How?

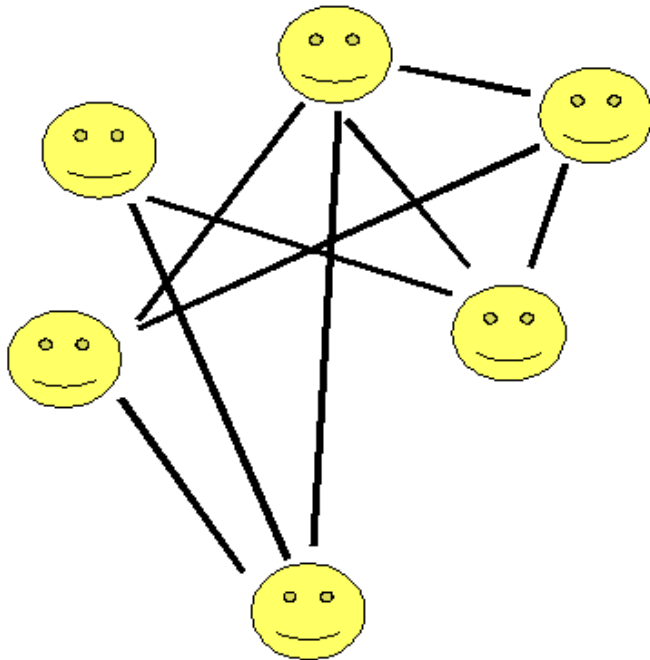
Open source collaboration tools are
good for you.

Open source collaboration tools



Sharing a vision

Mess media



No record of what was said.
Just one person in the know.

Central hub

[illegible]

Searchable, archived
Wikis, Mailinglists, blogs

What is the status - updates

- Fast feedback.



- Tools: svn, jira, mailinglists, continuum.

Real time help requests

- Track feature discussion.
- Pick problems to work on.

Assigned: 0 Unassigned: 1

Reporter: [Robert Burrell Donkin](#)

Votes: 0

Watchers: 0

Available Workflow Actions

☐ [Submit Patch](#)

☐ [Resolve Issue](#)

Operations

☐ [Assign this issue \(to me\)](#)

☐ [Attach file to this issue](#)

MAHOUT-156	Decode matrix methods	Unassigned	Daniel Nee	Open	UNRES
MAHOUT-80	Adding all scripts (for nightly build) to SVN repository	Unassigned	Edward J. Yoon <td>Open <th>UNRES</th> </td>	Open <th>UNRES</th>	UNRES
MAHOUT-82	Mahout-Hama integration	Unassigned	Edward J. Yoon <td>Open <th>UNRES</th> </td>	Open <th>UNRES</th>	UNRES

https://issues.apache.org/jira/browse/MAHOUT-85

Mahout

Add Element Labels to Vectors and Matrices

Created: 08 Jun 08 11:52 AM | Updated: Wednesday 01:51 PM

Component(s): [Matrix](#)

Affects Version(s): 0.1

Fix Version(s): None

Time Tracking: Not Spent

File Attachments:

(Click on column header to sort)

File Name	Size
MAHOUT-85-name_patch	20K
MAHOUT-85-name_patch	20K
MAHOUT-85-name_patch	20K
MAHOUT-85d_patch	20K
MAHOUT-85d_patch	20K
MAHOUT-85d_patch	20K
MAHOUT-85d_patch	20K

Issue Links:

Blocker

This issue blocks:

→ [MAHOUT-120](#) Prepare document vectors from the test

Description

Many applications can benefit by accessing elements in vectors and matrices using String labels in addition to:

[All](#) [Comments](#) [Watch Log](#) [Change History](#) [Subversion Commits](#) [Fetch Log](#)

[Jeff Eastman](#) added a comment - 08 Jun 08 12:06 PM

This patch introduces four new operations on Vectors:

1. `setLabelBindings(Map<String, Integer> bindings)` - sets a map of bindings between string labels and vector
2. `getLabelBindings()` - returns the map, creating an empty one if needed
3. `get(String label, double value)` throws `UnboundLabelException`, `IndexException` - sets the labeled element or
4. `get(String label)` throws `UnboundLabelException`, `IndexException` - returns the value indexed by the given label

This patch required reworking of the `serialize/deserialize` code so that the maps are included in the format run.

This is about as minimal a change as I could imagine, and I chose to just expose the map rather than try to do an kind of functionality that people will find useful?

[Jeff](#)

- Don't: "Just ask Bob to fix this."

Archives

- Documentation: No time.
- Traceability of code: SVN logs.
- Design discussions: JIRA.
- Searchable mailinglists.
- Rest: Wikis/blogs.



Mahout

A sub-project of Lucene

Luce





COMMUNITY NEWS

Finishing touches still to come

A glimpse of today, yesterday



M



January 3, 2006 by Matt Callow

<http://www.flickr.com/photos/blackcustard/81680010>

News aggregation



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

Ergebnisse. Seite 1 von 26 → in 3.451 sec

Telekom fordert Schadenersatz von Zumwinkel
Die Telekom fordert in der Spitzelaffäre Schadenersatz von Ex-Aufsichtsratschef Klaus Zumwinkel. - (© J. Hoffmann GmbH und Co. KG)
... so ein Telekom-Sprecher. Der Spiegel berichtete von den Schadenersatzansprüchen gegen Zumwinkel in Zusammenhang mit der Bespitzelungsaffäre ...
15:28 Uhr 18.04.2009 - [dieharte.de](#) - Politik

- [Telekom fordert Schadenersatz von Zumwinkel](#)
15:24 Uhr 18.04.2009 - [eiz.de](#) - Vermischtes
- [Telekom fordert Schadenersatz von Zumwinkel](#)
15:21 Uhr 18.04.2009 - [bb.live.de](#) - Vermischtes
- [Telekom fordert Schadenersatz von Zumwinkel](#)
15:14 Uhr 18.04.2009 - [wnoz](#) - Vermischtes

[Alle Suchergebnisse [zum Thema](#) - mehr als 66 Nachrichten]

Die Opfer der Telekom-Panne
Wer zahlt für entstandene Schäden? 20 bis 30 Millionen Kunden waren stundenlang nicht per Handy erreichbar. - (© Zeitungsverlag Ruhrgebiet Gmt & Co)
Ein Softwareproblem bei der Telekom hat den bislang größten Ausfall im deutschen Mobilfunknetz verursacht. 20 bis 30 Millionen Kunden waren mehrere ...
22:42 Uhr 22.04.2009 - [WE waz](#) - Vermischtes

Telekom senkt Prognose für 2009
Die Deutsche Telekom hat ihre Erwartungen für das laufende Jahr zurückgenommen. - (© Deutsche Presse-Agentur GmbH)
Sie rechnet nun mit einem Rückgang des Gewinns um Zinsen, Steuern und

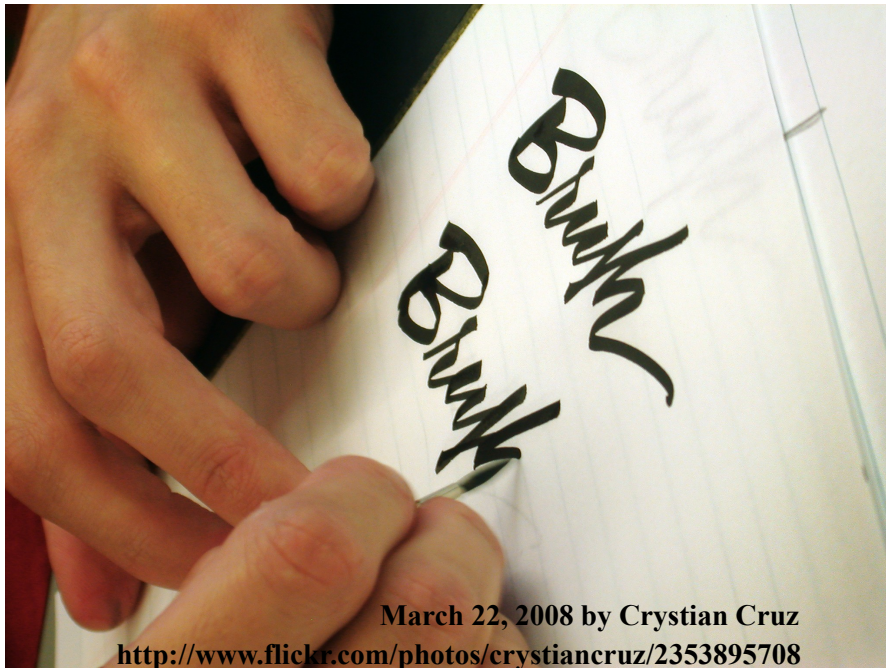
Wish: Aggregate sources
and track emerging topics.

17 30
20 45
SO 14 30
BRAD PITT DIANE KRUGER
INGLOURIOUS BASTERDS.
BY QUENTIN TARANTINO
DOLBY STEREO SR-D

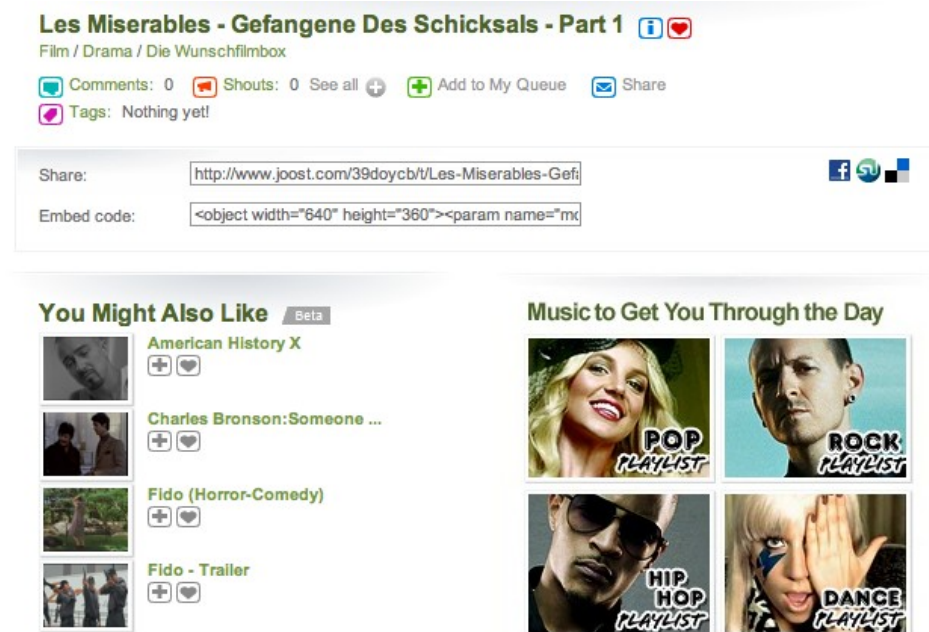
ODEON



Go to cinema



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



Wish: Reviews, sentiment detection, recommendations.

Machine learning – what's that?

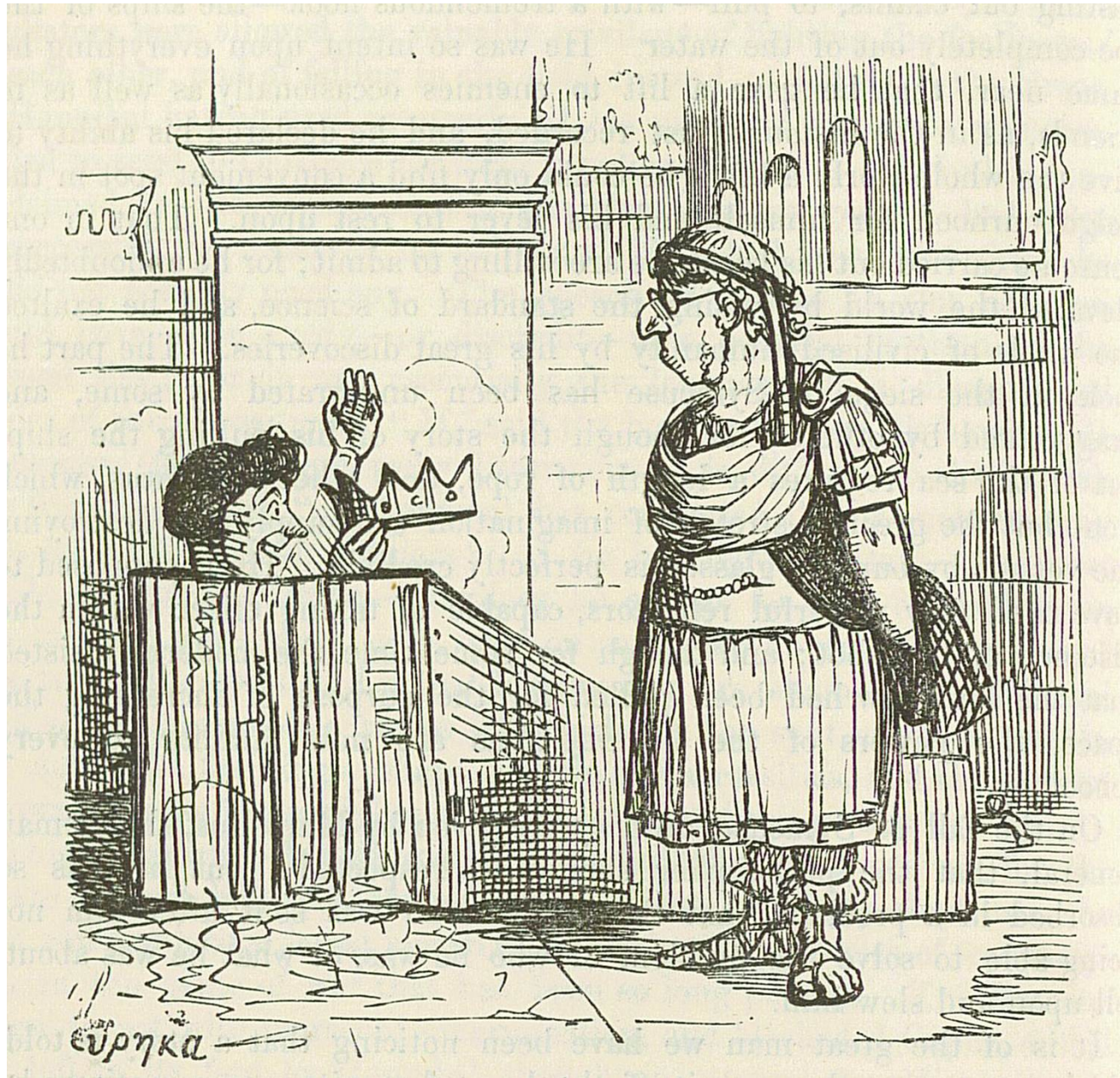


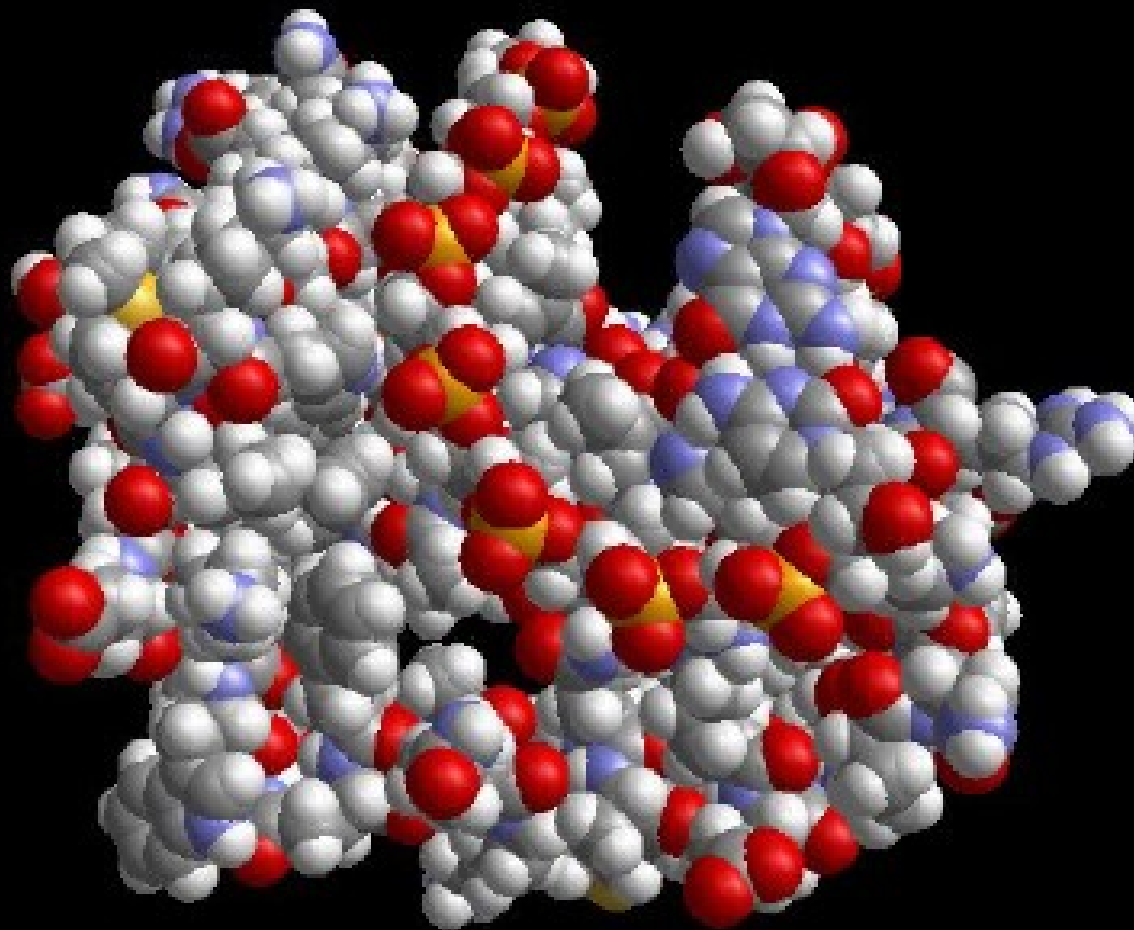
Image by John Leech, from: *The Comic History of Rome* by Gilbert Abbott A Beckett.
Bradbury, Evans & Co, London, 1850s
Archimedes taking a Warm Bath

Archimedes model of nature

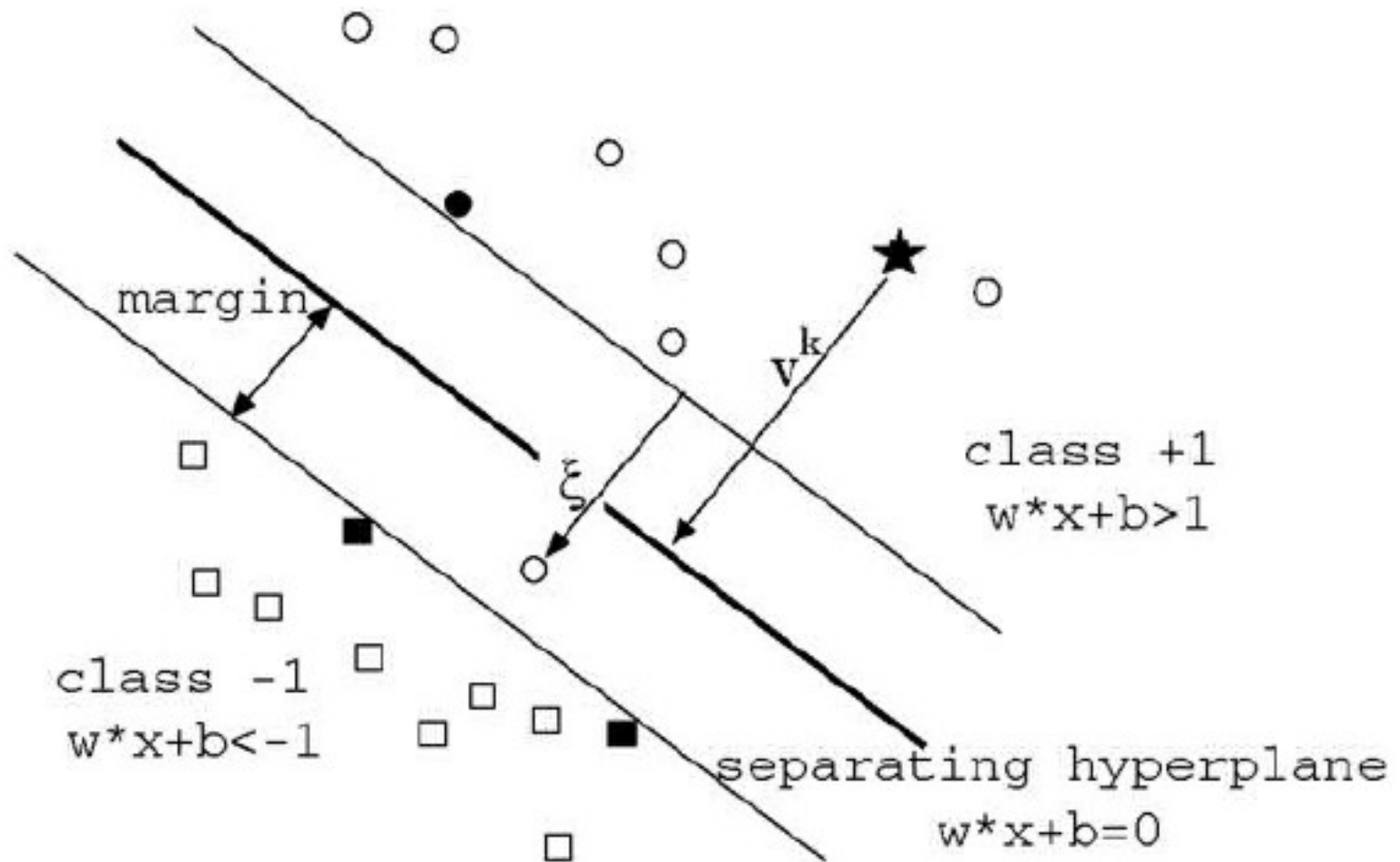
$$\frac{\textit{Density of Object}}{\textit{Density of Fluid}} = .$$

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





An SVM's model of nature



The challenge

- Large amounts of data.
- Structured and unstructured data.
- Diverse tasks.

Mission

Provide scalable data mining algorithms.

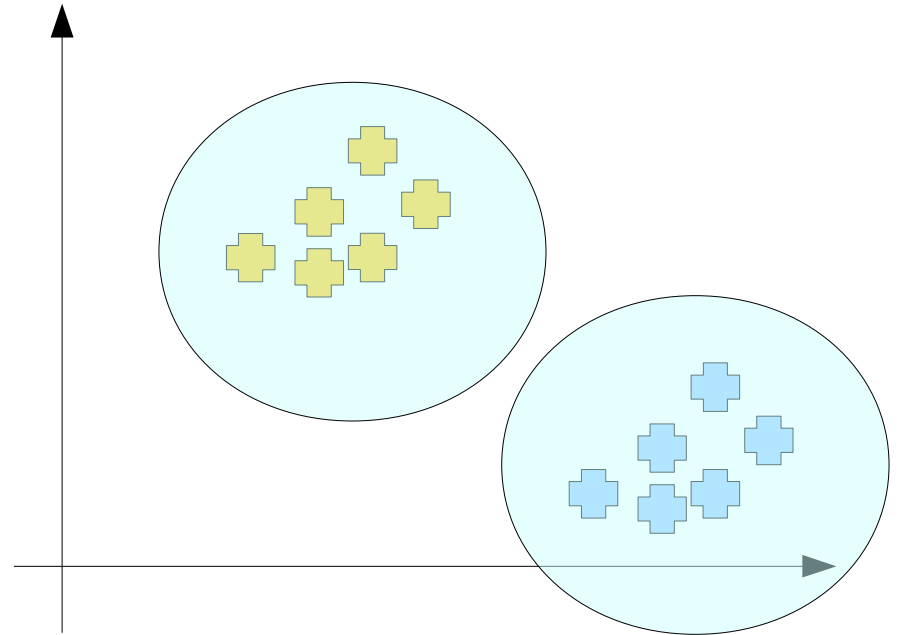


- Commercially friendly license.
- Scalable to large amounts of data.
- Well documented.
- Healthy community.
- Targeted to developers.

What does Mahout have to offer.

Discover groups of items

- Group items by similarity.



- Examples:
 - Group news articles by topic.
 - Find developers with similar interests.

U.K.

World

[Top Stories](#)

[World](#)

[U.K.](#)

[Business](#)

[Sci/Tech](#)

[Entertainment](#)

[Sports](#)

[Health](#)

[Spotlight](#)

[Most Popular](#)



Telegraph.co.uk

[Qaeda-linked group claims Baghdad bomb attacks](#)

Reuters - [Andrew Hammond](#) - 2 hours ago

DUBAI (Reuters) - An al Qaeda-linked group has said it carried out the twin suicide bombings that killed 155 people in Baghdad on Sunday and revived doubts about security in the run-up to Iraq's elections in January.

[Video: Too early for US to withdraw from Iraq](#) RT

[Al-Qaida linked group claims Baghdad attacks](#) The Associated Press

[Aljazeera.net](#) - [BBC News](#) - [Sky News](#) - [Washington Post](#) - [Wikipedia: 25 October 2009](#)

[Baghdad bombings](#)

[all 3,834 news articles »](#) [Email this story](#)



Times Online

[Obama vows no rush on Afghanistan](#)

BBC News - 3 hours ago

US President Barack Obama has said he will "never rush" a decision to send more troops to Afghanistan, as he comes under pressure to set out a new policy.

[Video: Obama resists pressure on Afghan war strategy - 27 Oct 09](#) Al Jazeera

[Obama refuses to rush troops decision](#) ABC Online

[New York Times](#) - [Reuters India](#) - [The Associated Press](#) - [AFP](#)

[all 1,665 news articles »](#) [Email this story](#)



Times Online

[Karadzic court case due to resume](#)

BBC News - 1 hour ago

The genocide and war crimes trial of former Bosnian Serb leader Radovan Karadzic is due to resume in The Hague, a day after it was adjourned.

[Video: Karadzic is a surrogate Milosevic in The Hague](#) RT

[Karadzic snubs his war crimes trial..but it will go ahead without him](#) Mirror.co.uk


[guardian.co.uk](#) - [New York Times](#) - [The Associated Press](#) - [Independent](#)

[all 1,214 news articles »](#) [Email this story](#)

[All news](#)

[Headlines](#)

[Images](#)



[Web](#) [MSN](#) [Yahoo](#) [Wiki](#) [Images](#) [News](#) [J](#)

mahout

Tree Visualization

All Topics (100)

Machine Learning (10)

Mahout Project (10)

Thai Mahout (7)

Introducing Apache Mahout (6)

Scalable Machine Learning (6)

Tests (6)

Laos (5)



Working (5)

ApacheCon (4)

Day Mahout Training (4)




more | show all

Top 100 results of about 35900 for ma

1 Apache Mahout - Overview  
Mahout's goal is to build scalable,
<http://lucene.apache.org/mahout/>

2 Mahout - Wikipedia, the free encyc
A mahout is a person who drives a
<http://en.wikipedia.org/wiki/Mahout>

3 mahout - Definition from the Merri
Function: noun. Etymology: Hindi &
<http://www.merriam-webster.com/d>

4 What is a Mahout?   
Brief and Straightforward Guide: WI
<http://www.wisegeek.com/what-is-a>

Discover groups of similar items

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

Discover groups of similar items

- Example: Synthetic Control
 - <http://archive.ics.uci.edu/ml/datasets/Synthetic+Control>
 - Example Job: `<MAHOUT_HOME>/examples`
 - Outputs clusters
- Download the distribution.
- Run the example.
- Have a closer look at the examples.

Identify dominant topics

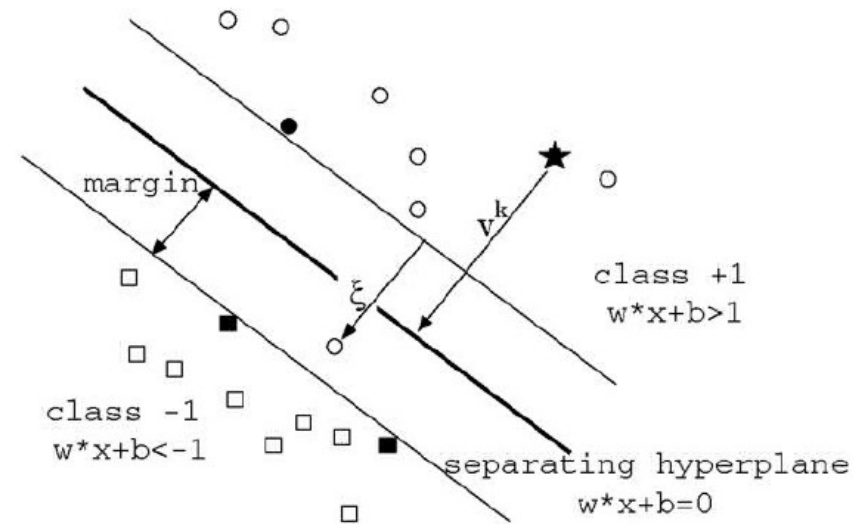
- Given a dataset of texts, identify main topics.

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.



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



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

[Advanced Image Search](#)

 SafeSearch: [Moderate](#) ▼

 Images ☐ [Hide options](#)

Results

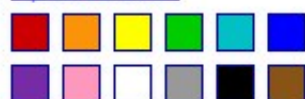
> Any size

[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](#)

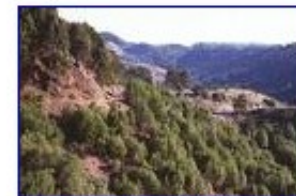
The Oakland Baseball

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](#)

oakland

Search images

[Advanced Image Search](#)

SafeSearch: [Moderate](#) ▼

[Images](#) > [Face](#) ☐ [Hide options](#)

Results 1 - 20 of about 1,400,000 (0.1)

> Any size

[Medium](#)

[Large](#)

[Icon](#)

[Larger than...](#)

[Exactly...](#)

Any type

[Face](#)

[Photo](#)

[Clip art](#)

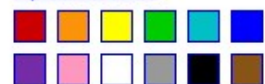
[Line drawing](#)

> 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](#)



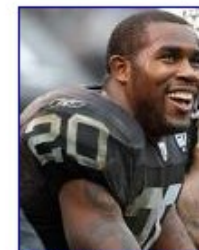
Dallas Cowboys v
594 x 404 - 60k
[zimbio.com](#)



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.
- Others upcoming.

Assign items to defined categories

- Examples based on “standard” datasets:

- 20 Newsgroups

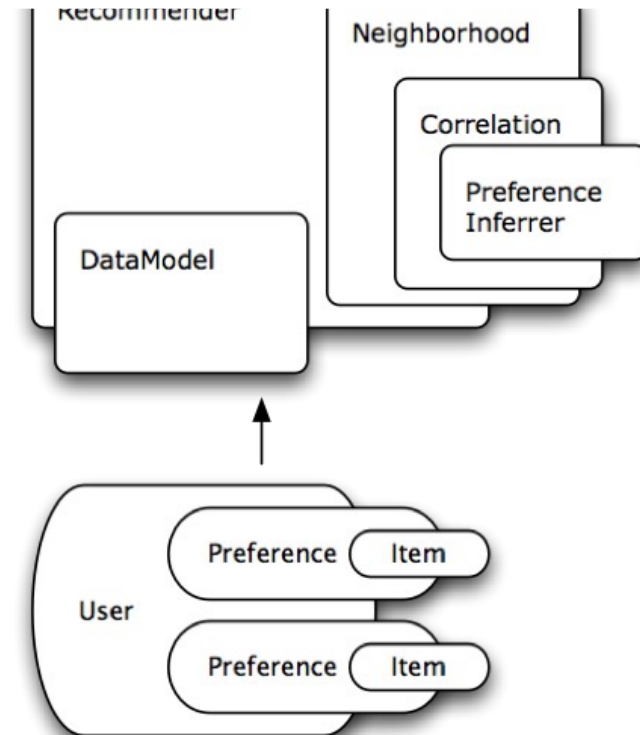
<http://cwiki.apache.org/MAHOUT/twentynewsgroups.html>

- Wikipedia

<http://cwiki.apache.org/MAHOUT/wikipediabayesexample.html>

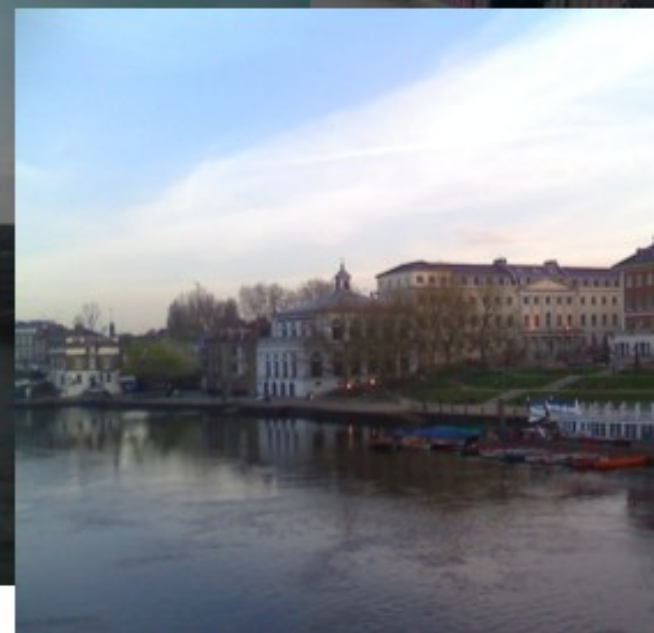
Recommendation mining.

- Recommend items to users.



- Examples:
 - Find books related to the book I am buying.
 - Find movies I might like.

Recommending places



Recommending people

People you may know

Shane Curcuru
Conference Lead at Apache
Softw

Ugo
Com
Profe

Arjé
CTO
lead,
Mem
Foun

Invite | x

Invite | x

Invite | x



Recommendation mining.

- Integrated Taste.
- Mature Java library.
- Java-based, web service / HTTP bindings.
- Batch mode based on EC2 and Hadoop.

Frequent pattern mining

- Given groups of items, find commonly co-occurring items.
- Examples:
 - In shopping carts find items bought together.
 - In query logs find queries issued in one session.





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/>



Upcoming

- More algorithms.
- Optimization of existing implementations.
- More examples.
- Release 0.3

Jumpstart your project with proven code.



January 8, 2008 by dreizeh
<http://www.flickr.com/photos/1328/2176949>

A photograph of two men in a dimly lit room. The man on the left is wearing a grey hoodie and has his hand to his face, looking towards the other man. The man on the right is wearing a black t-shirt and glasses, sitting in a chair with his hand on his chin, looking back at the first man. The background is a plain, light-colored wall.

Discuss ideas and problems online.

November 16, 2005 [p
<http://www.flickr.com/photos/hi-phi/6405>



Become part of the community.



mahout-user@lucene.apache.org

mahout-dev@lucene.apache.org



Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.

Image by: Patrick McEvoy

Isabel Drost
Jan Lehnardt
newthinking store
Simon Willnauer



BERLIN BUZZWORDS 2010

This is to announce the Berlin Buzzwords 2010 scalability conference. Berlin Buzzwords 2010 is scheduled for the start of June. Topics of interest include NoSQL databases, Hadoop, Lucene and others. Our goal is to bring developers and users together in central Europe for a conference featuring talks on scaling data analysis. The [team](#) organizing this event is deeply rooted in the Hadoop, Lucene, and CouchDB communities. Interested in helping? See the [requests for helping hands](#). Also note that we are just getting off the ground. Please be patient as we get the various infrastructure pieces in place.

June 7/8th: Berlin Buzzwords 2010

Store, Search, Scale

BERLIN BUZZWORDS NEWS
FEBRUARY 2009 - CFP TO BE PUBLISHED

The call for presentations will be published on this site in mid-February (including more detailed

Solr

HBase

Lucene

Sphinx

Hadoop

Distributed computing

CouchDB

Business Intelligence

NoSQL

Cloud Computing

Scalability

MongoDB

Mar., 10th 2010: Hadoop* Get Together in Berlin

- Bob Schulze (eCircle/ Munich): Database and Table Design Tips with HBase
- Dragan Milosevic (zanox/ Berlin): Product Search and Reporting powered by Hadoop
- Chris Male (JTeam/ Amsterdam): Spatial Search

Apache Hadoop Get Together Berlin March 2010

Wednesday March 10, 2010 at 5:00pm

[newthinking store](#)

Tucholskystr. 48

Berlin, Bundesland Berlin [Get Directions](#)

Event Photos



[+ Add Photos](#)

[See al](#)

<http://upcoming.yahoo.com/event/5280014/>

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

mahout-user@lucene.apache.org

mahout-dev@lucene.apache.org



Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.

Image by: Patrick McEvoy

Why?

Why should I waste my time with
doing stuff for free?

Work on what you want...

when you want.



<http://www.flickr.com/photos/abnelgonzalez/2058764760/>

Share and discuss with peers.



Learn from the best.



<http://www.flickr.com/photos/mg315/381296439/>

Soft Skills.



<http://www.flickr.com/photos/ajawin/3587215356/>

Make work visible and re-usable.



Get started

Turn users into developers.

GSoC

ComDev

How to Contribute to Mahout

"Contributing" to an Apache project is about more than just writing code – it's about doing what you can to make the project better. There are lots of ways

- [How to Contribute to Mahout](#)
 - [Be Involved](#)
 - [Contributing Code \(Features, Big Fixes, Tests, etc...\)](#)
 - [Getting the source code](#)
 - [Making Changes](#)
 - [Generating a patch](#)
 - [Unit Tests](#)
 - [Creating the patch file](#)
 - [Contributing your work](#)
 - [Review/Improve Existing Patches](#)
 - [Applying a patch](#)
 - [Helpful Resources](#)

Be Involved

Contributors should join the [Mahout mailing lists](#). In particular:

- the user list (to help others)
- The commit list (to see changes as they are made)
- The dev list (to join discussions of changes)

Please keep discussions about Mahout on list so that everyone benefits. Emailing individual committers with questions about specific Mahout issues is discouraged. See http://people.apache.org/~hossman/#private_q.

Contributing Code (Features, Big Fixes, Tests, etc...)

This section identifies the "optimal" steps community member can take to submit a changes or additions to the Mahout code base. This can be new features or changes to existing code to prove it works as advertised (and to make it more robust against possible future changes).

Please note that these are the "optimal" steps, and community members that don't have the time or resources to do everything outlined on this below should follow "Yonik Seeley's (Solr committer) Law of Patches" ...

HowToBecomeACommitter

Added by [Grant Ingersoll](#), last edited by [Grant Ingersoll](#) on Apr 13, 2008 ([view change](#))

How To Become A Committer

While there's no exact criteria for becoming a committer, there is a fairly obvious path to becoming a committer.

For starters, one should be familiar with the [Apache Way](#), especially the part about meritocracy.

Second, participate in the mailing lists, help answer questions when you can and do so in a respectful manner. This is often more important than writing amazing code.

Third, write code, add patches, stick with them and be patient. Add unit tests and documentation. In general, tackling 3 or 4 decent patches is where the bar is at, but in the early stages of the project, the bar is a bit lower, so it pays to join early!

Finally, it is then up to someone to nominate them to the PMC. Typically, one of the existing committers does this by sending an email to the private PMC mailing list and then the PMC votes on it. Nominations often occur internal to the PMC as well.

mahout-user@lucene.apache.org

mahout-dev@lucene.apache.org



Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.

Image by: Patrick McEvoy