

Scale your data application with
Apache Mahout





Isabel Drost

NOKIA

Daytime:

Co-Founder Berlin Buzzwords.
Software developer at Nokia Gate 5 GmbH.

Nighttime:

Co-Founder Apache Mahout.
Member Apache Software Foundation.
Founder of Berlin Hadoop Get Together.

CONTRA-INDICATIONS DOSAGE SIDE-EFFECTS PREGNANCY OVERDOSE
IDENTIFICATION PATIENT INFORMATION



faker PLACEBO forte™[®] Tablets
Code: 0815/4711
SCHEDULING STATUS:
Not scheduled
PROPRIETARY NAME
(and dosage form):
faker PLACEBO forte™[®] Tablets

COMPOSITION:
300 mg **Nihilit** (virtually nothing) per tablet
PHARMACOLOGICAL CLASSIFICATION:
A 0.0 Psychological and Imaginary Antianxiety
PHARMACOLOGICAL ACTION:
faker PLACEBO forte™[®] has no pharmacological actions whatsoever.

CONTRA-INDICATIONS:
None.
WARNINGS:
If symptoms persist, you should increase the dose. Do not use for less than 10 days.

DOSAGE AND DIRECTIONS FOR USE:
Adults: 1 to 3 bubble packs. Repeat every five minutes if necessary but not less than 12 tablets to be taken during any 24 hour period.

SIDE EFFECTS AND SPECIAL PRECAUTIONS:
None. This is REALLY safe!
Interference with laboratory tests:
Nihilit occasionally explodes unexpectedly.

KNOWN SYMPTOMS OF OVERDOSAGE AND PARTICULARS OF ITS TREATMENT:
None. In cases of overdosage, just take another faker PLACEBO forte™[®] to counter the side-effects.
PRESENTATION:
300 mg tablets in blister packs of 100, 50, 30, 10.

STORAGE INSTRUCTIONS:
Don't store. Just take them all as quickly as possible and buy a new package. Let the kids have some, too.

REGISTRATION NUMBER:
m/3 1/41457
NAME AND BUSINESS ADDRESS OF APPLICANT:
faker (Pty) Ltd, 23 Foo Road, Barbown, 1300. 42/23/13

DATE OF PUBLICATION OF THIS PACKAGE INSERT:
December 2012
faker = Revoked Fake Mark of Fakes AG, Germany

PATIENT INFORMATION LEAFLET
Ignore this leaflet before you start treatment. If any of this information causes you to just skip this and take another pill.
Composition:
Each tablet contains Nihilit 300 mg. The medicine also contains nothing else.



Hello GeeCon visitors.

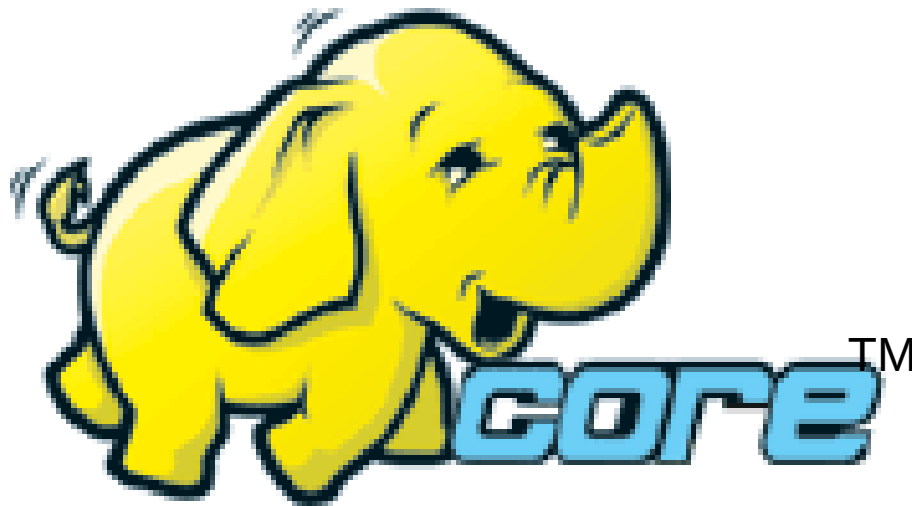
Apache Hadoop

Know it?

Use it?

How many nodes?

Contributed patches?



Apache Lucene/ Solr



Apache Mahout





Image thanks to Philipp Kaden and Berlin Buzzwords
<http://flickr.com/photomicc>
<http://flickr.com/berlinbuzzwords>

Data-based approach



Image by Leo Reynolds:
<http://www.flickr.com/photos/lwr/3552240620>

Find requirements from past users' behaviour



Users seldom give explicit feedback.

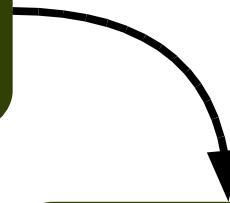
Take a snapshot of user interactions.

Analyze what they do.

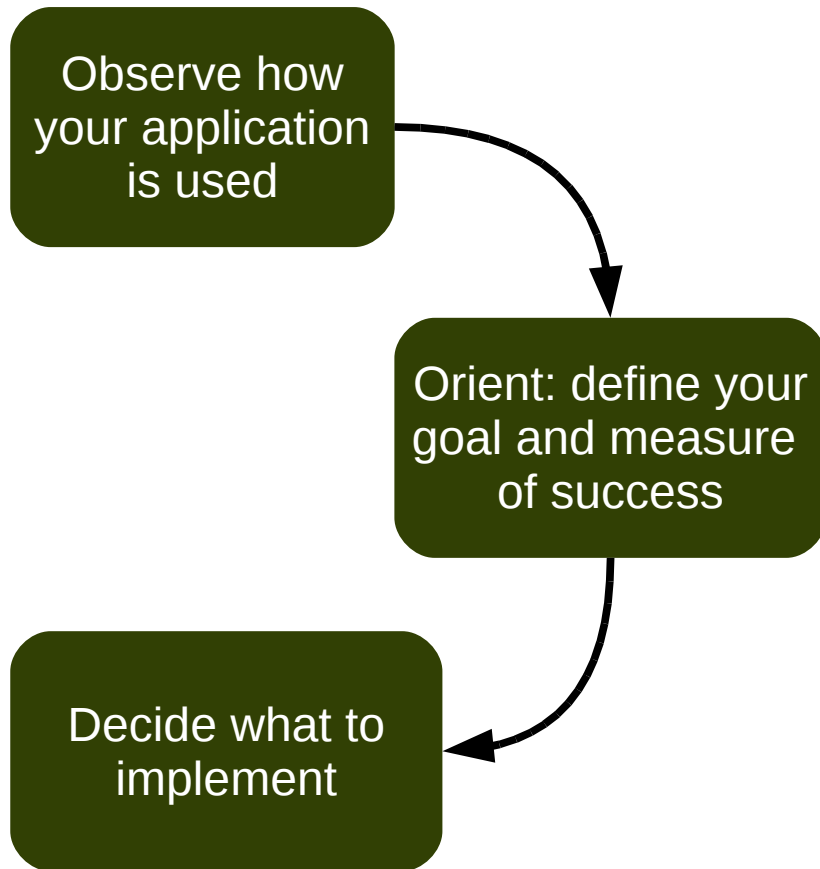
Derive needs for new functionality from that.

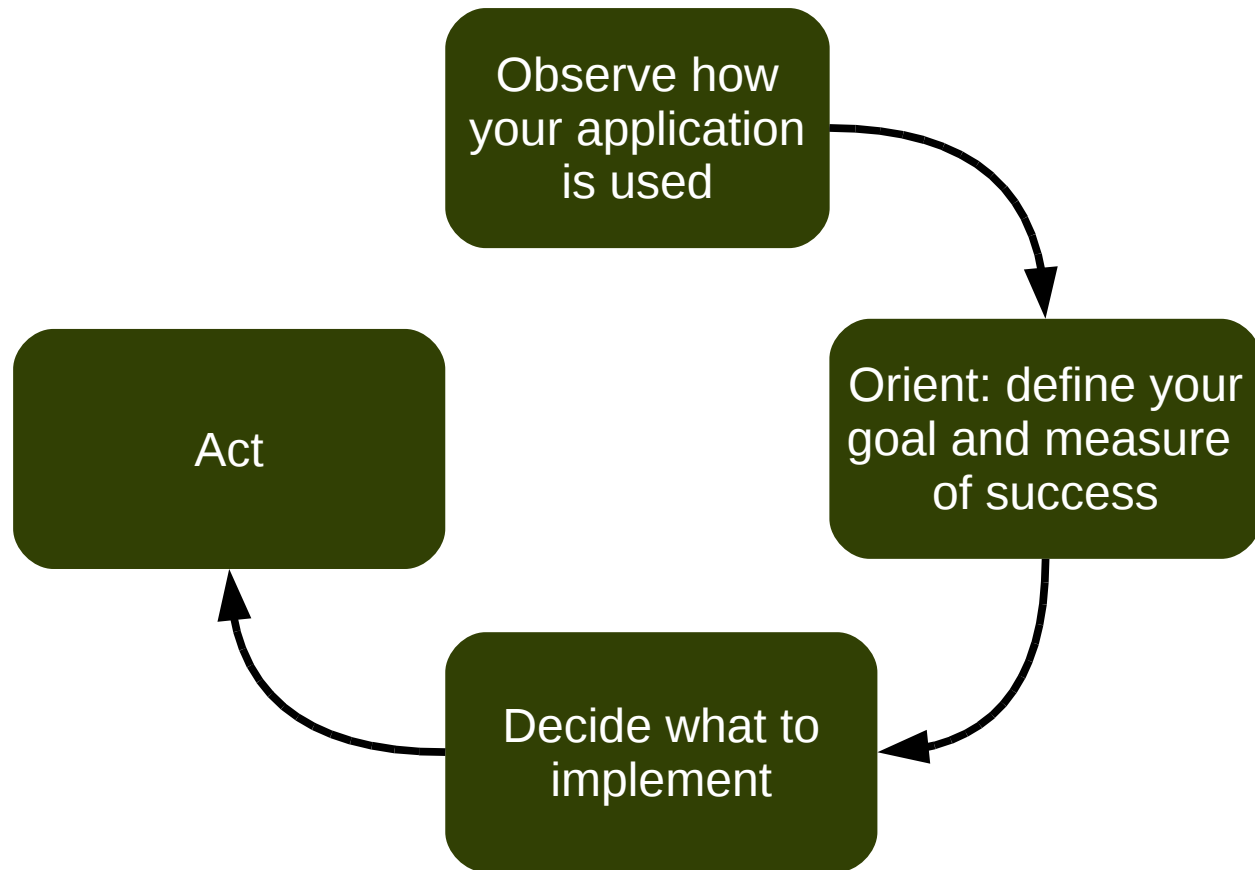
Observe how
your application
is used

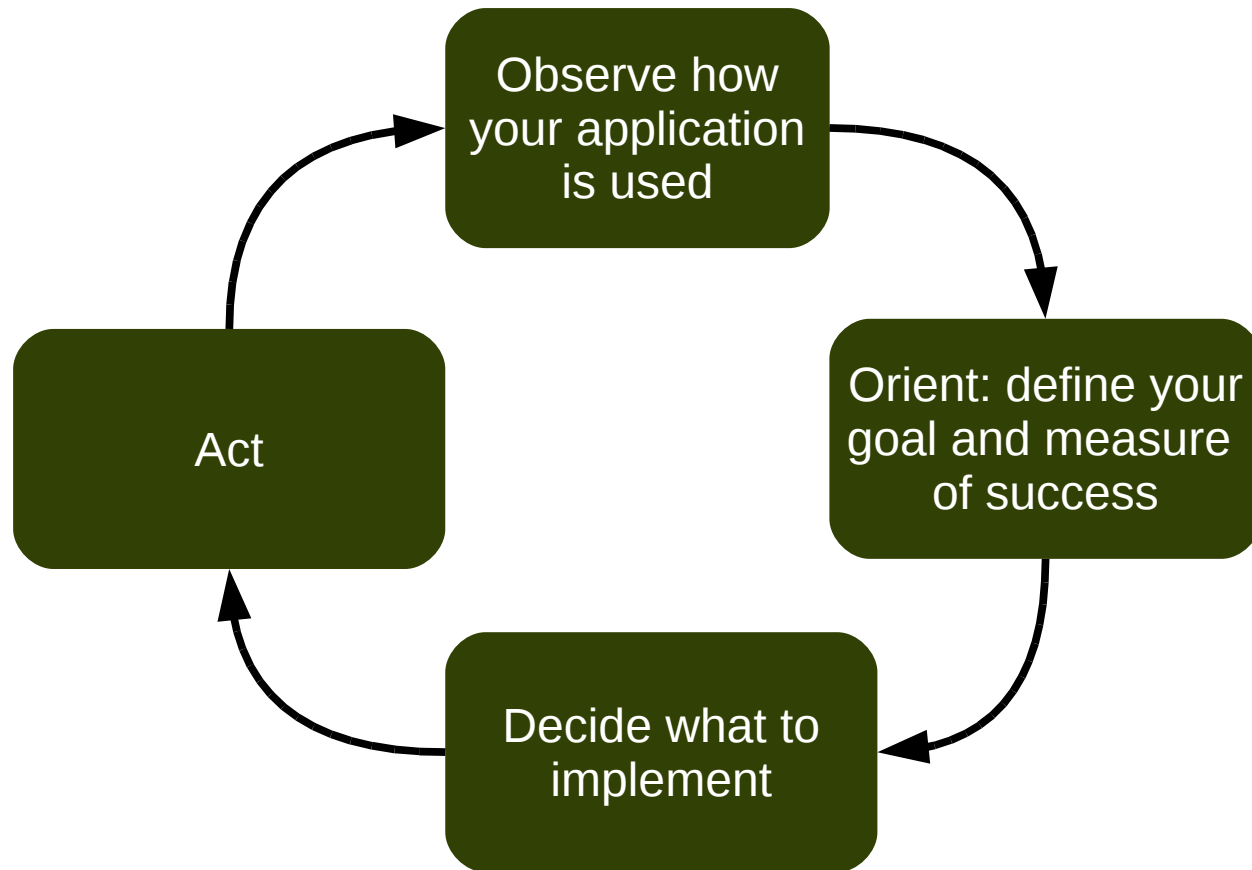
Observe how
your application
is used

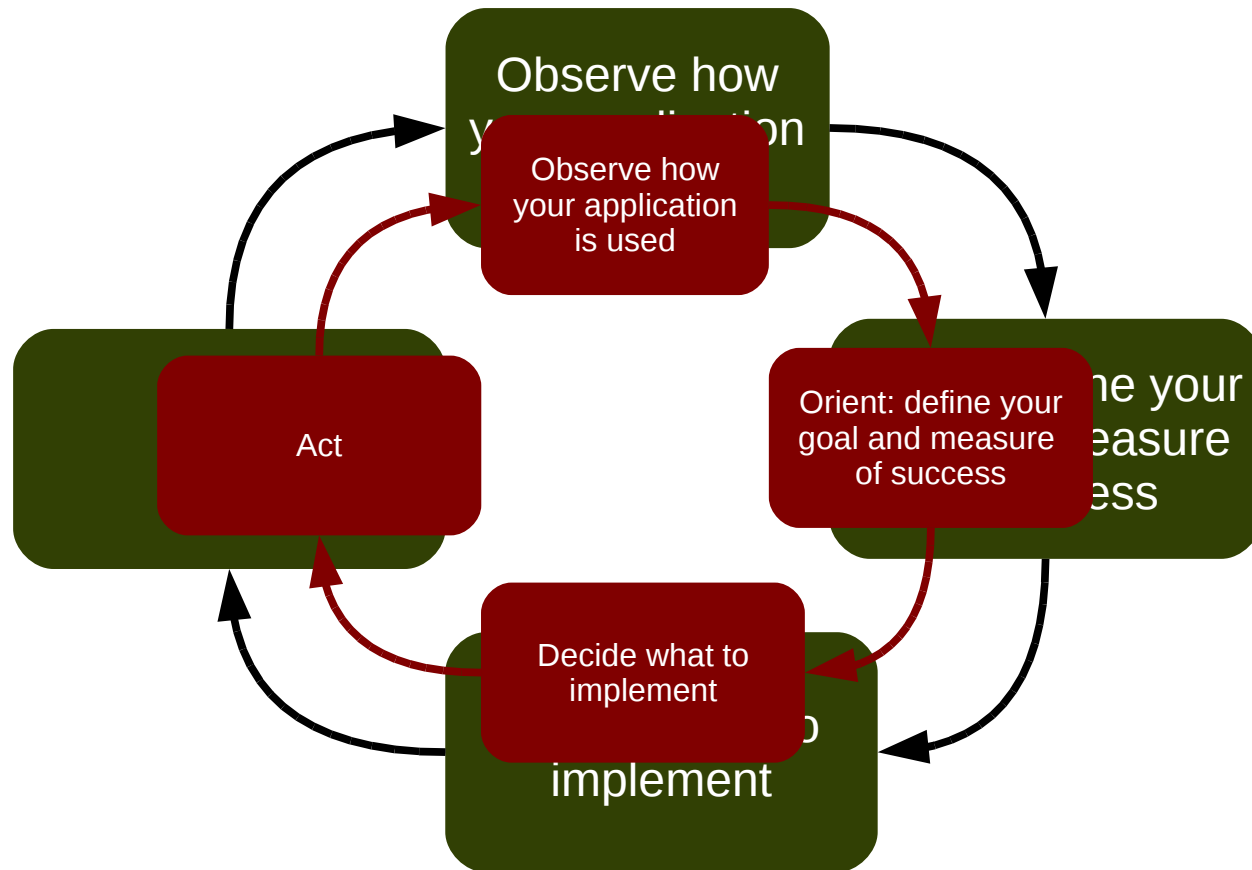


Orient: define your
goal and measure
of success









Develop new features that rely on user input



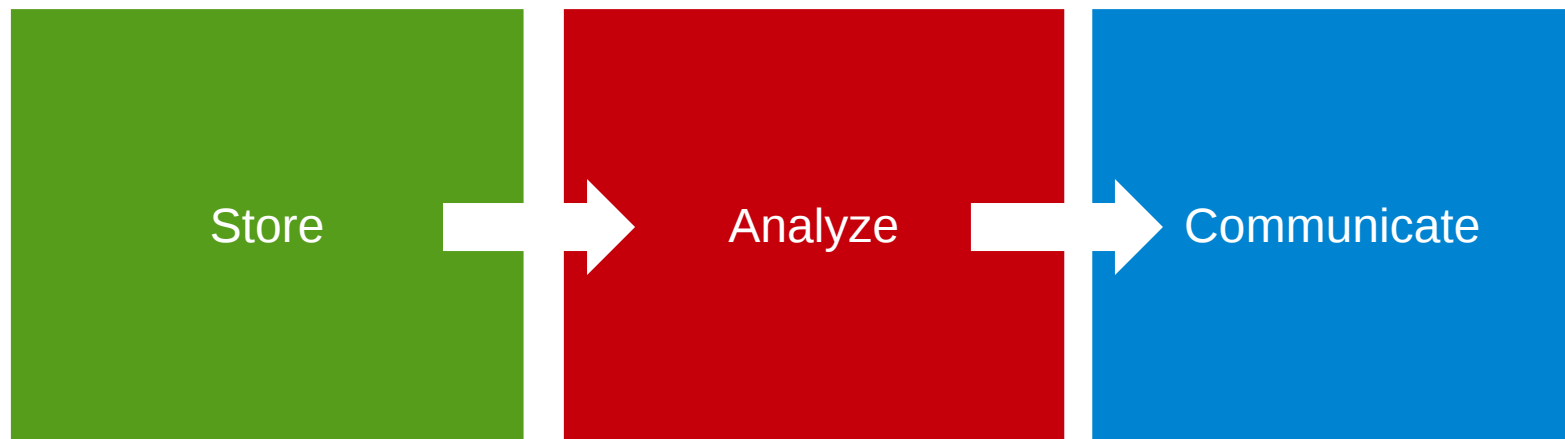
Recommend places based on my friends' preferences

Show content on site based on what I clicked and read

Automatically identify the important mails in my inbox

Discover new blogs based on what I read

From grouping and counting
to machine learning






```
2 03 00 Ending seckpack...
printBuffer(): g_hlen = 12, g_data =
02 00 05 02 02 00 00 00 00 00 00 00
printBuffer(): g_hlen = 12, g_data =
02 00 05 00 02 00 00 00 00 00 00 00
Erasing to finish... g_data = 0
printBuffer(): g_hlen = 16, g_data = 0
Waiting for...
printBuffer(): g_hlen = 16, g_data = 0
02 00 06 00 00 00 00 00 00 00 00 00
08 03 00 Enjoy your unlocked iPhone...
Unmounting filesystems...
Checking HFS Plus volume.
/dev/rdisk0s1
Checking a case-sensitive catalog.
Checking Extended Attributes file.
** Detected a case-sensitive file.
** Checking Catalog hierarchy.
** Checking Catalog Attributes.
** Checking Extended Attributes.
** Checking volume information.
** Checking volume LittleBear-H102..UserBundle
** Checking volume OK.
** The volume is a case-sensitive catalog.
** appears to be HFS Plus volume.
** /dev/rdisk0s2
** Checking a case-sensitive file.
** Detected a case-sensitive file.
** Checking Catalog file.
** Checking Catalog file.
```

Extracting information from log files

- Access logs.
- Health check results.
- Response time logs.

New requirements:

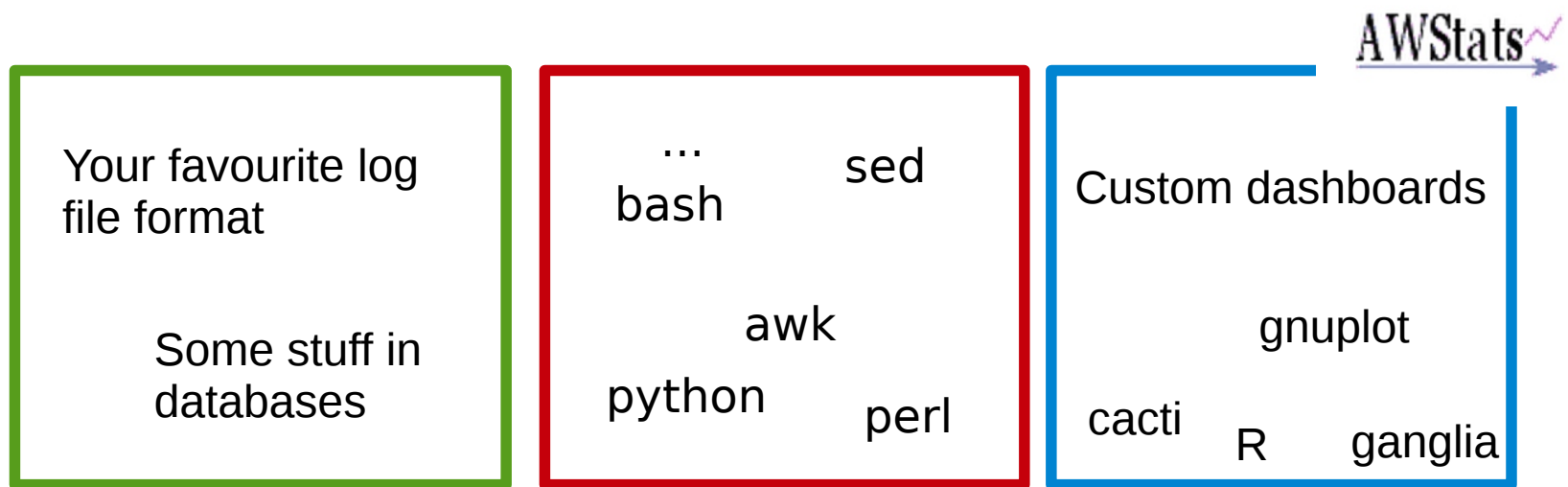
- Predict traffic for machine sizing
- Harden setup against intrusion attempts
- Analyze outages for future prevention

Data based features:

- Show only content I have not seen yet
- Show different content based on region
- Show content I like based on click stream

Extracting information from log files

- Access logs.
- Health check results.
- Response time logs.





Structured relational data

- Attendee database.
- Speaker information.
- User transactions.

New requirements:

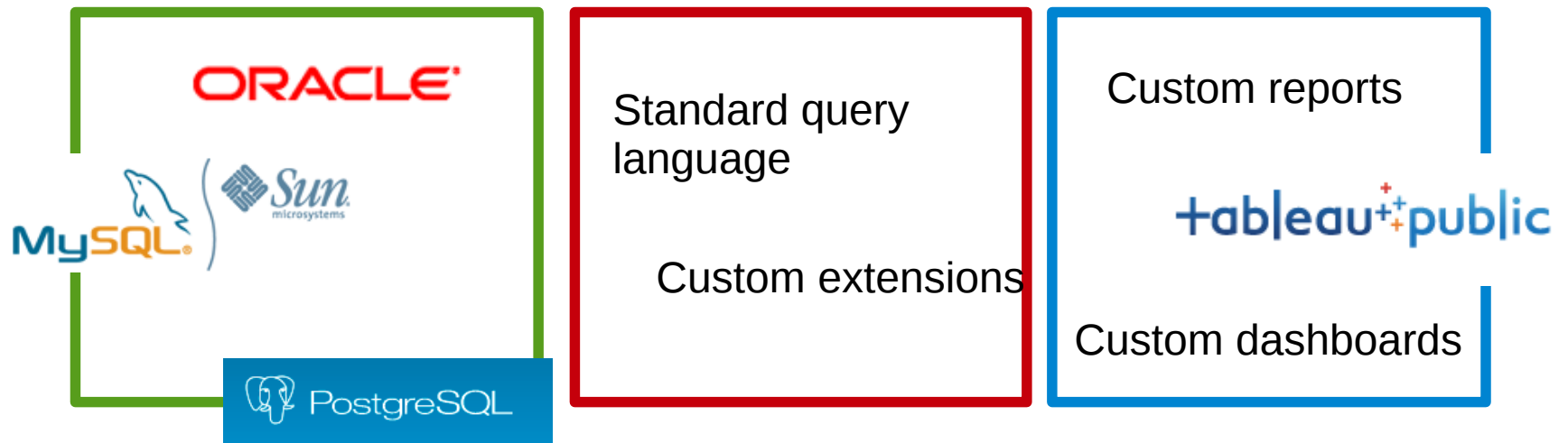
- Missing audience.
- Define lunch menu.
- Travel budget.

Data based features:

- Show selected content based on age.
- Recommendation of talks based on explicit topics of interest.

Structured relational data

- Attendee database.
- Speaker information.
- User transactions.





Google code **protobuf**
Protocol Buffers - Google's data interchange format



Java API
Pipes C++ API
Streaming
Dumbo for Python
...



Cascading

Export much smaller
result set to database
and use standard tools.

Use custom visualization.

Export to file format
compatible with your
favourite graphing tool.

What if...



Nutrition Facts
Valeur nutritive
Per 1/2 can (125 g) (1/2 can 1/2 can 1/2 can)
Amount
Calories 180
Fat 10g
Sodium 100mg
Total Crap 10g
Protein 10g
Fiber 10g
Sugar 10g
Calcium 10g
Iron 10g

Everything

Images

Maps

Videos

News

Shopping

More

Any time

Past 24 hours

Past week

Custom range...

All results

By subject

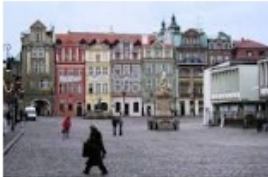
Any size

Large

Medium

Icon

Related searches: [poznan poland](#) [poznan map](#) [poznan malta](#) [poznan university of medical sciences](#)



Maps

Videos

News

Shopping

More

Any time

Past 24 hours

Past week

Custom range...

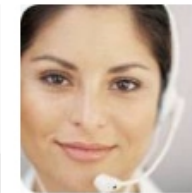
All results

By subject

Any size

Large

Medium



images

Maps

Videos

News

Shopping

More

Any time

Past 24 hours

Past week

Custom range...

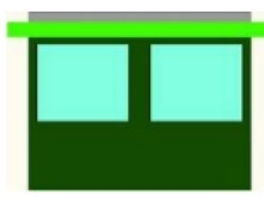
All results

By subject

Any size

Large

Related searches: [poznan poland](#) [poznan map](#) [poznan malta](#) [poznan university of medical sciences](#)



POZnan*



**Chaos CC**

Auch aktueller Staatstrojaner rechtswidrig

Dem Chaos Computer Club (CCC) ist eine «noch fast fabrikneue» Version des umstrittenen Staatstrojaners zugespielt worden.

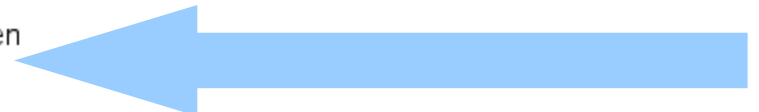
Berlin (dpa) - Dem Chaos Computer Club (CCC) ist eine «noch fast fabrikneue» Version des umstrittenen Staatstrojaners zugespielt worden. Eine Analyse des Programmcodes habe ergeben, dass die Spionage-Software weiterhin für rechtswidrige Aktionen eingesetzt werden könne, erklärte der Club am Mittwoch.

Anzeige

Wo die Software eingesetzt wurde, blieb zunächst offen. Das Bundesinnenministerium dementierte den Einsatz für die ihm unterstellten Behörden, das sind Bundeskriminalamt (BKA), Bundesverfassungsschutz und Bundespolizei. Vor zweieinhalb

Schlagwörter

Wochen hatte der CCC eine erste Trojaner-Software zum Abhören
BKA Bundeskriminalamt Bundespolizei Bundesverfassungsschutz CCC Chaos
Computer Club Innenausschuss Kommunikation Trojaner



Top Stories

News near you

World

Revolutionary Armed
Forces of Colombia

Daniel Ortega

Syria

Benazir Bhutto

Bangkok

Boko Haram

Antonis Samaras

Gaza

Suicide attack

U.S.

Business

Technology

Entertainment

Sports

Science

Health

Spotlight

World



Financial ...

Greek PM Calls For Meeting With Opposition Chief, President on Coalition

Voice of America - 11 minutes ago



November 06, 2011 Greek PM Calls For Meeting With Opposition Chief, President on Coalition VOA News Greek Prime Minister George Papandreou says he wants to meet with the country's president and opposition leader in the coming hours to agree on a new ...

[Euro debt crisis: Greek PM George Papandreou to resign when new coalition ...](#) Telegraph.co.uk

[Europe turns up heat on Greeks to seal unity coalition | Reuters](#)

From Greece: [Samaras: If the PM resigns, I am determined to help](#) Kathimerini

In Depth: [Pressure Mounts on Greek Premier to Resign](#) New York Times

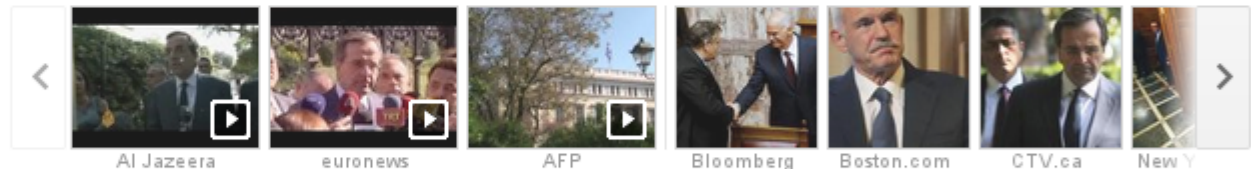
[See all 889 sources »](#)

Related

[Antonis Samaras »](#)

[George Papandreou »](#)

[Karlolos Papoulias »](#)



Globe and ...

Syrian forces kill 13 as Arab League calls meeting

Jerusalem Post - 52 minutes ago

Analyst tells 'Post' "Arab League has intentions to support uprising in Syria but needs US commitment to support them if Iran retaliates.



Reuters Afri...

Nigerian forces hunt killers

Reuters - 50 minutes ago

KANO (Reuters) - Nigerian security forces said on Sunday they were searching for Islamist militants behind a coordinated attack in the north that killed at least 65 people, as residents still in shock demanded the government do more to protect them.

Computer

 Sortiert nach: **Standard**


- Standard
- Top-Quellen
- Aktualität
- Relevanz

Kein Hacker-Angriff auf Facebook

N24.DE VOR 55 MINUTEN

Die Hacker-Gruppen Anonymous hat keinen Angriff auf Facebook gestartet. Sprecher der Hacker-Gruppe erklärten, dass ein isoliertes Mitglied dazu aufgerufen habe und nicht unterstützt worden sei.


Meldungen (133), **Kommentare** (0)


Apple unterliegt Motorola vor Gericht

HARTWARE VOR 22 STUNDEN

[ZUM ARTIKEL](#) 

Motorola Mobility konnte sich vor dem Landgericht Mannheim gegen Apple durchsetzen. Das Gericht sah es als erwiesen an, dass Apple zwei Motorola-Patente verletzt, und verhängte ein Verkaufsverbot für mobile Geräte. Das ...


Meldungen (86), **Kommentare** (0)



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: W
http://www.wisegeek.com/what-is-a

Show most relevant products

clearasil®

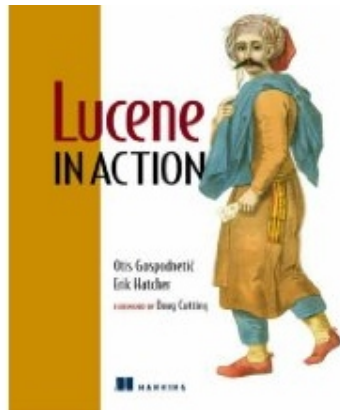
erwandte Suchbegriffe: clearasil pickelstift, world of warcraft.

este Ergebnisse

« Zurück | Seit

1.  **Clearasil 44161 Tiefenreinigung Antibakterielle Reinigungspads, 60er Packung**
Neu kaufen: **EUR 5,99** 2 Angebote ab **EUR 5,15**
Lieferung bis **Samstag, 22. März**: Bestellen Sie innerhalb der nächsten **23 Stunden**
Kostenlose Lieferung möglich.
★★★★★ (1)
Drogerie & Bad: Alle 13 Artikel ansehen
2.  **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
3.  **World of Warcraft: Wrath of the Lich King (Add-on)** von Vivendi Unive
Vista / XP)
Neu kaufen: **EUR 39,99**
Vorbestellbar
Kostenlose Lieferung möglich.
Games: Alle Artikel ansehen

Show most relevant products



~ [Otis Gospodnetic](#) (Author), [Erik Hatcher](#) (Author)
★★★★☆ (20 customer reviews)

List Price: ~~\$44.95~~

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

In Stock.

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

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

Also Available in: List Price: Our Price: Other Offers:

Paperback (2) ~~\$44.99~~ **\$29.69**

[See larger image](#)

[Share your own customer images](#)

[Publisher: learn how customers can search inside this book.](#)

Frequently Bought Together

Customers buy this book with [Building Search Applications: Lucene, LingPipe, and Gate](#) by Manu Konchady



+



Price For Both: \$70.12



Add both to Cart

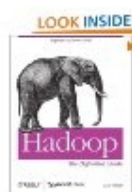
Add both to Wish List

[Show availability and shipping details](#)

Customers Who Bought This Item Also Bought



[Solr 1.4 Enterprise Search Server](#) by David



[Hadoop: The Definitive Guide](#) by Tom White



[Hibernate Search in Action](#) by Emmanuel



[Introduction to Information Retrieval](#) by



[Collective Intelligence in Action](#) by Satnam Alag



Recommending 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



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



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

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



Problems

- Often unstructured, non-relational data.
- No simple, well defined rules for prediction.
- Pattern search too tedious for humans.

Machine learning – what's that?

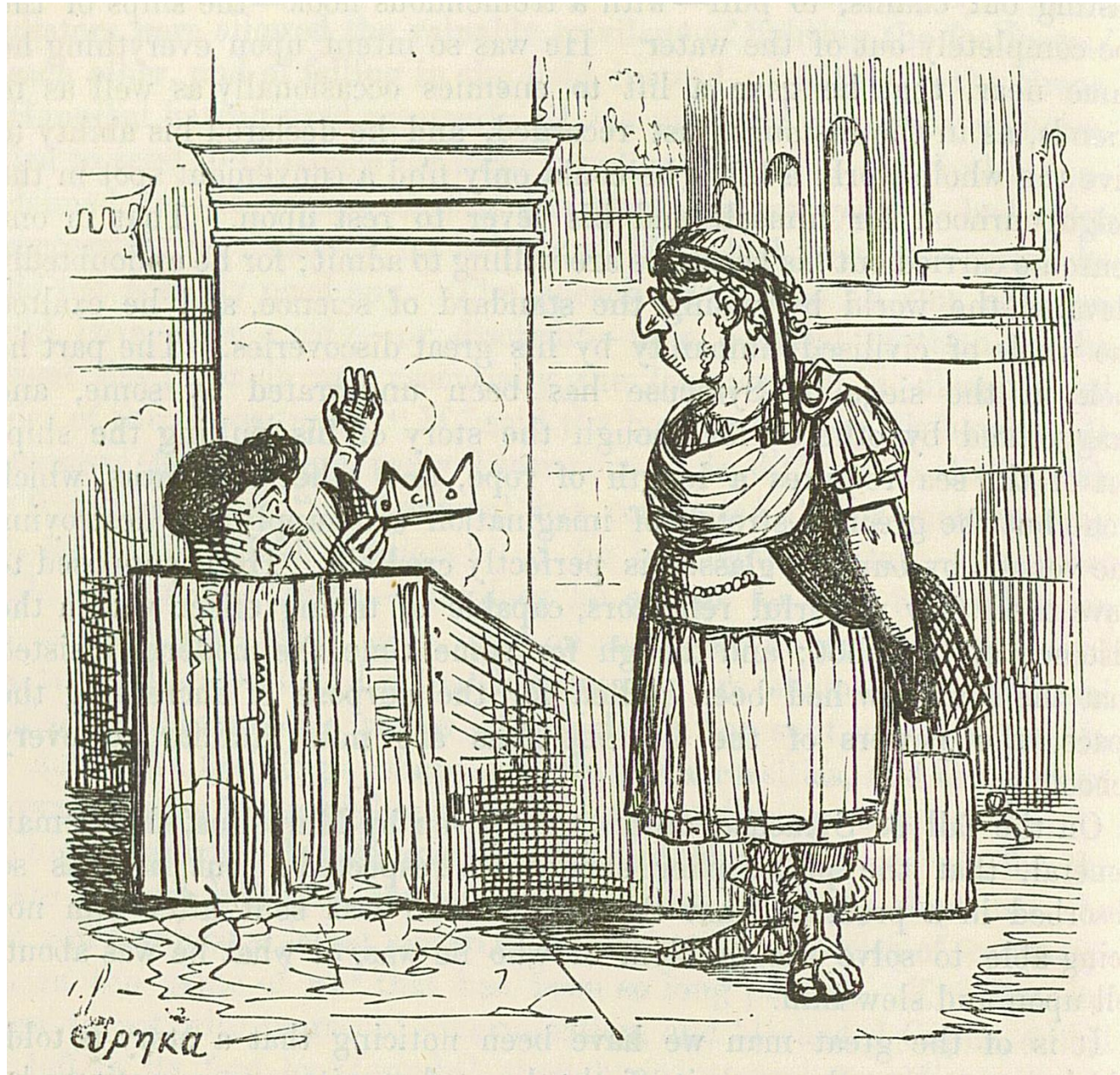


Image by John Leech, from: *The Comic History of Rome* by Gilbert Abbott A Beckett.
Bradbury, Evans & Co, London, 1850s
Archimedes taking a Warm Bath

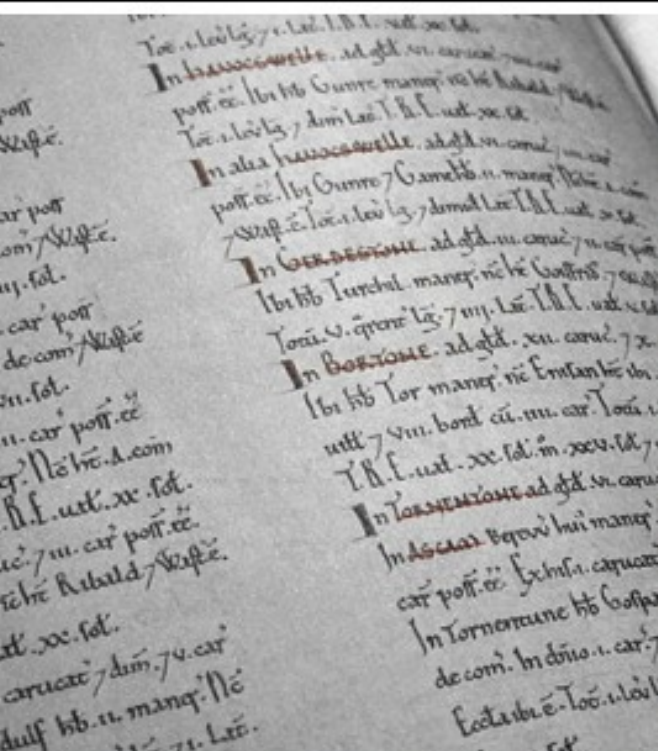
Archimedes model of nature

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

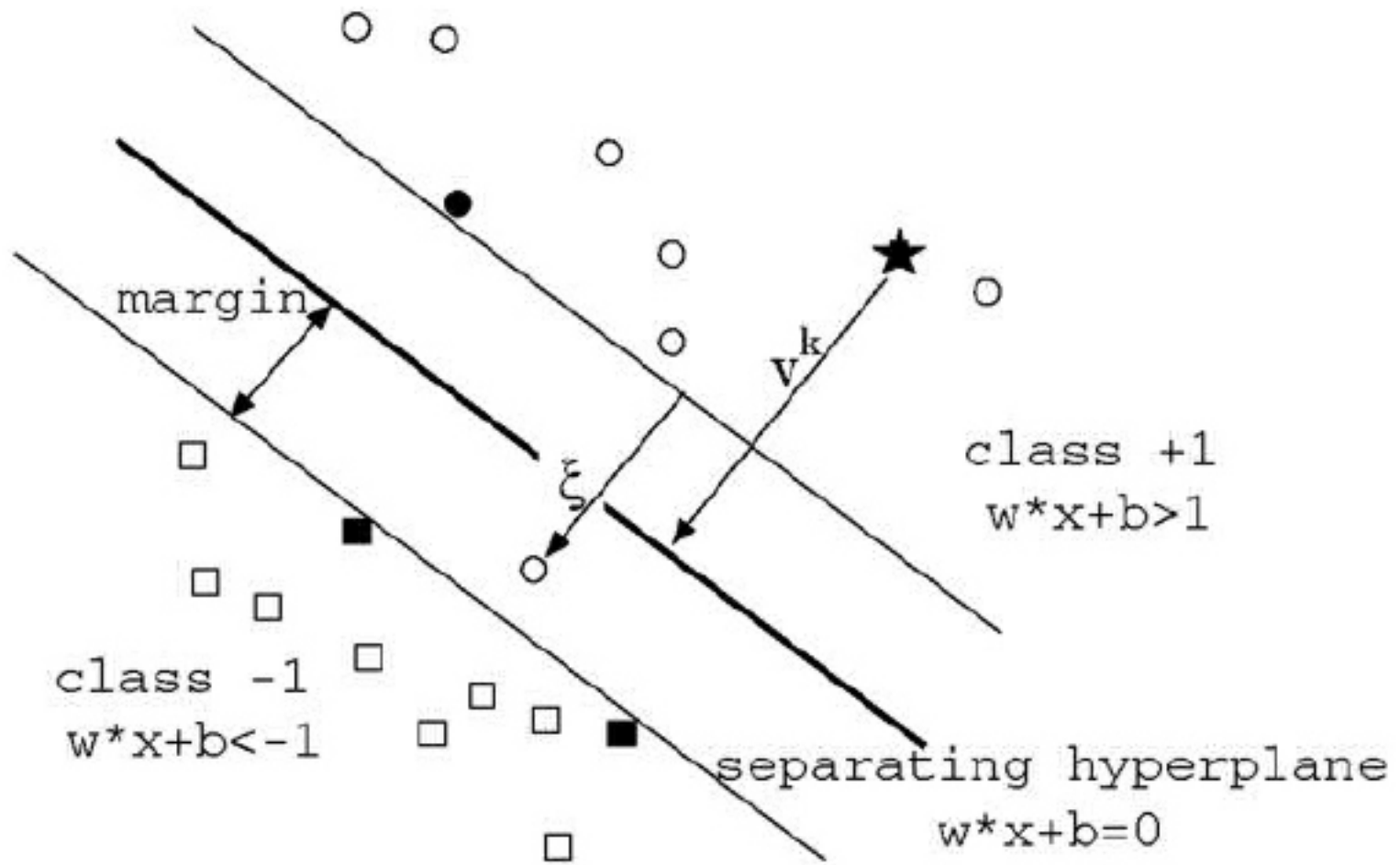
$$\frac{\textit{Weight}}{\textit{Weight} - \textit{Apparent immersed weight}}$$



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



An SVM's model of nature



The challenge

Mission

Provide scalable data mining algorithms.





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
And many more

Community



Emeritus:
Niranjan Balasubramanian
Erik Hatcher
Ozgur Yilmazel
Dawid Weiss



Image by Joe Thorn: <http://www.flickr.com/photos/joethorn/242599072>



Image by curtm95: <http://www.flickr.com/photos/curtm95/4599105577>



Image by leafnode: <http://www.flickr.com/photos/7881771@N07/3422736557>



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



Pieces in a machine learning project

Model

Instance*

(sometimes also called example, or item)

Feature*

(sometimes also called attribute or signal)

Label*

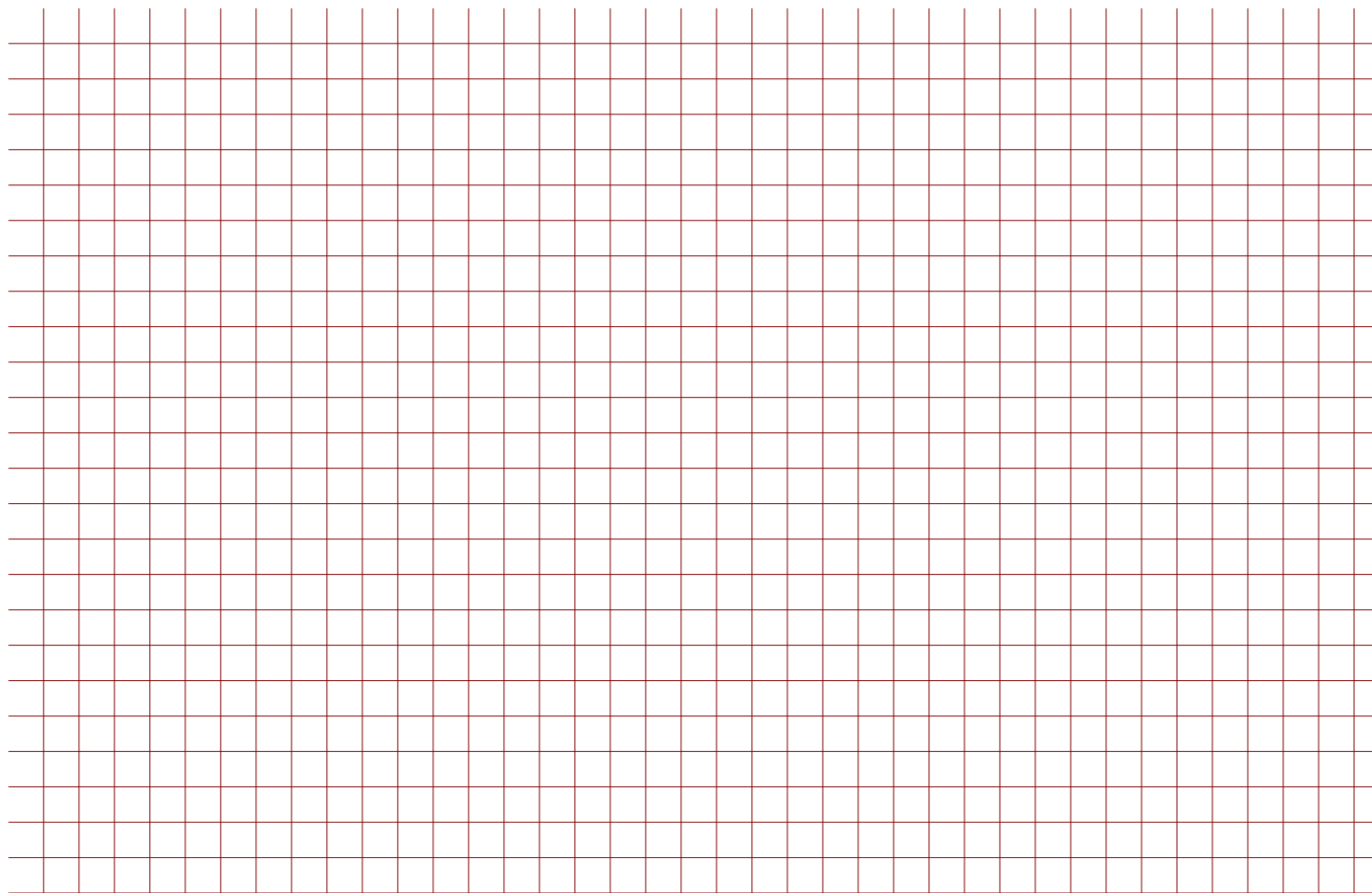
(sometimes also called class, target variable)




[illegible][illegible]

[illegible][illegible][illegible]

Learning problems in context





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 ...
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: W
http://www.wisegeek.com/what-is-a

World

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

Booked Knowledge

all 3,834 news articles > Email this story

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 RT

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

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

[illegible][illegible]This image shows a full page of graph paper. The background is a solid light blue color. Overlaid on this background is a uniform grid of thin, dark red lines. The grid consists of small squares that cover the entire area of the page, leaving no margins or other markings.



By freezeflight, <http://www.flickr.com/photos/68056612@N00/155554663/>

Google images

oakland

Search images

Advanced Image Search

SafeSearch: [Moderate](#) ▼

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

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

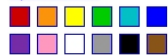
> Any size

[Medium](#)
[Large](#)
[Icon](#)
[Larger than...](#)
[Exactly...](#)

[Any color](#)
[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



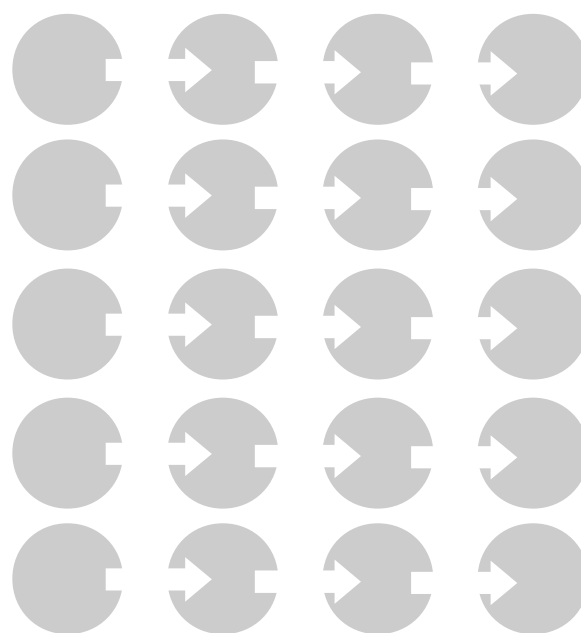
Distributed by Tubemogul. The
720 x 480 - 18k
channels.com



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



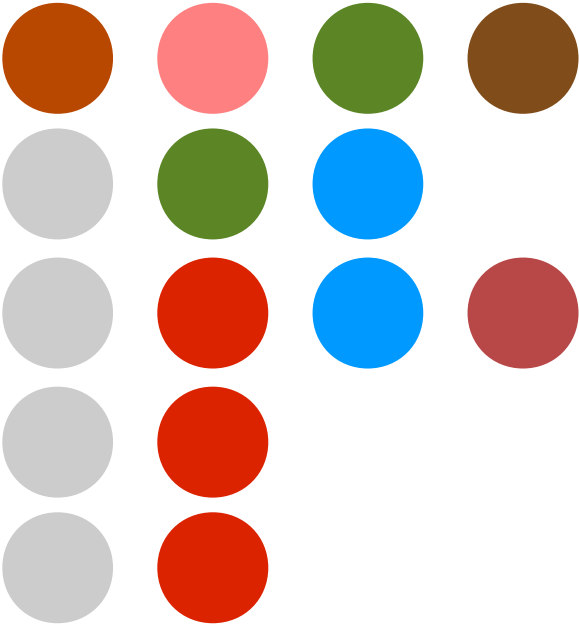
#20 of the Oakland
467 x 594 - 86k
zimbio.com





<http://www.flickr.com/photos/disowned/1158260369/>

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**.

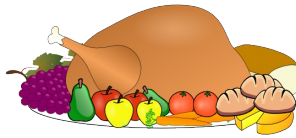


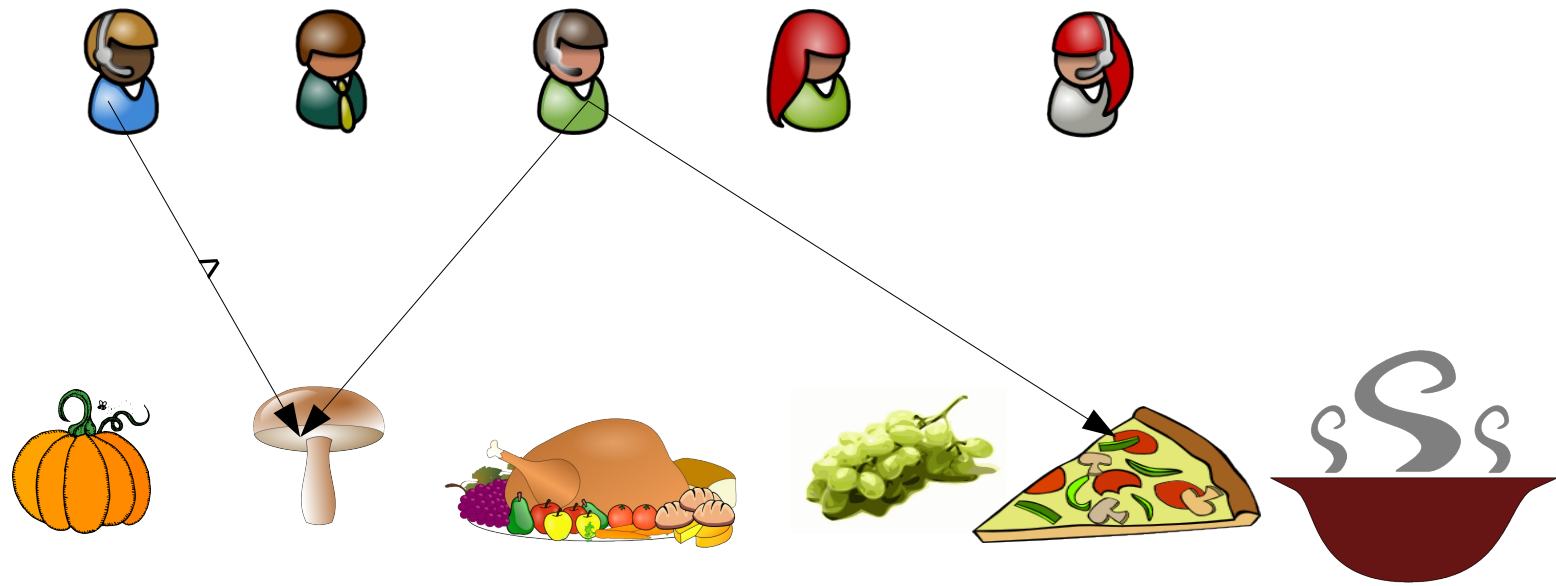


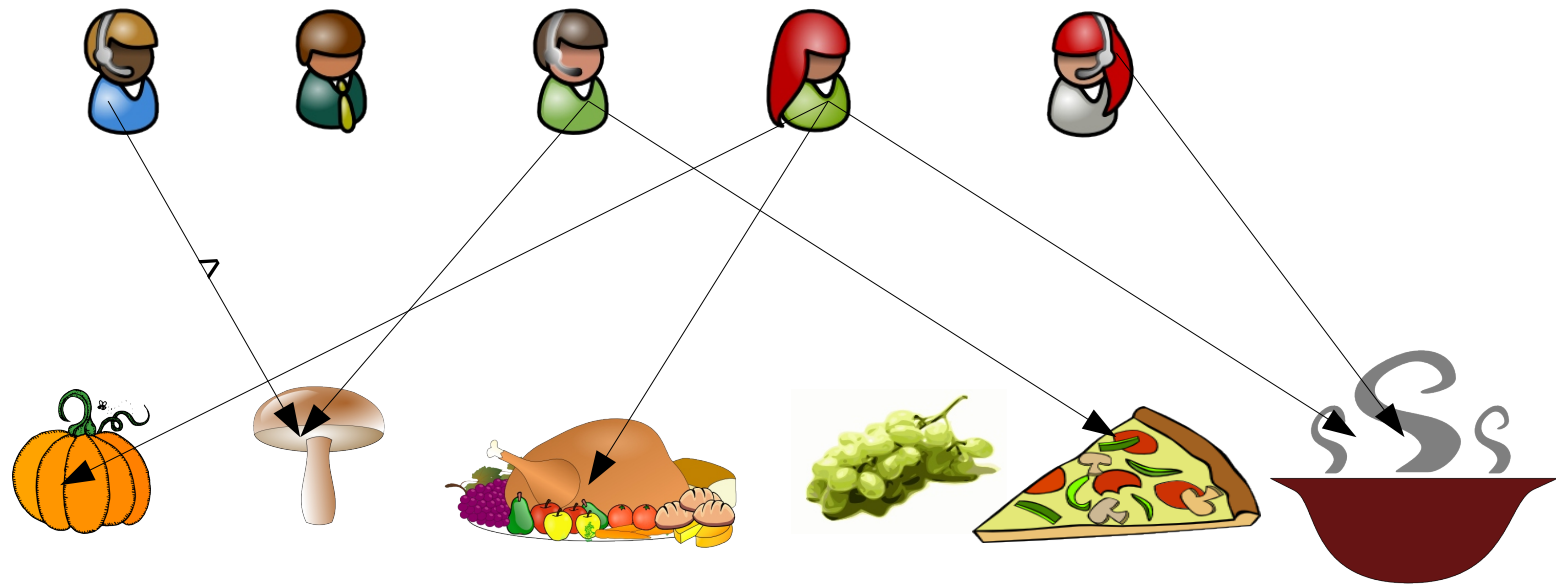
By quinnanya, <http://www.flickr.com/photos/quinnanya/2806883231/>








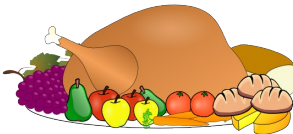

By crypto, <http://www.flickr.com/photos/crypto/3201254932/sizes//>

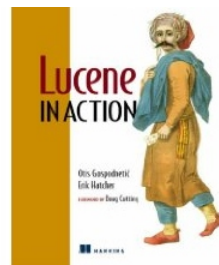
By libraryman, <http://www.flickr.com/photos/libraryman/78337046/sizes//>









[See larger image](#)

[Share your own customer images](#)

[Publisher: learn how customers can search inside this book.](#)

~ [Otis Gospodnetic](#) (Author), [Erik Hatcher](#) (Author)
★★★★☆ (20 customer reviews)

List Price: ~~\$44.95~~

Price: **\$29.67** & this item ships for **FREE with Super Saver Shipping.** [Details](#)

You Save: **\$15.28 (34%)**

In Stock.

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

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

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

Frequently Bought Together

Customers buy this book with [Building Search Applications: Lucene, LingPipe, and Gate](#) by Manu Konchady



Price For Both: \$70.12

[Add both to Cart](#)

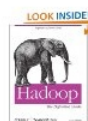
[Add both to Wish List](#)

[Show availability and shipping details](#)

Customers Who Bought This Item Also Bought



[Solr 1.4 Enterprise Search Server](#) by David



[Hadoop: The Definitive Guide](#) by Tom White



[Hibernate Search in Action](#) by Emmanuel



[Introduction to Information Retrieval](#) by



[Collective Intelligence in Action](#) by Satnam Ahuja

Your typical machine learning project



GREAT PARTY!
THANKS GUYS!
LOVE & HUGS

mmm.
lovely.
fruity

POLY 45

chalky

BADGERZ

Pump that bass!!!

Wicked Party!

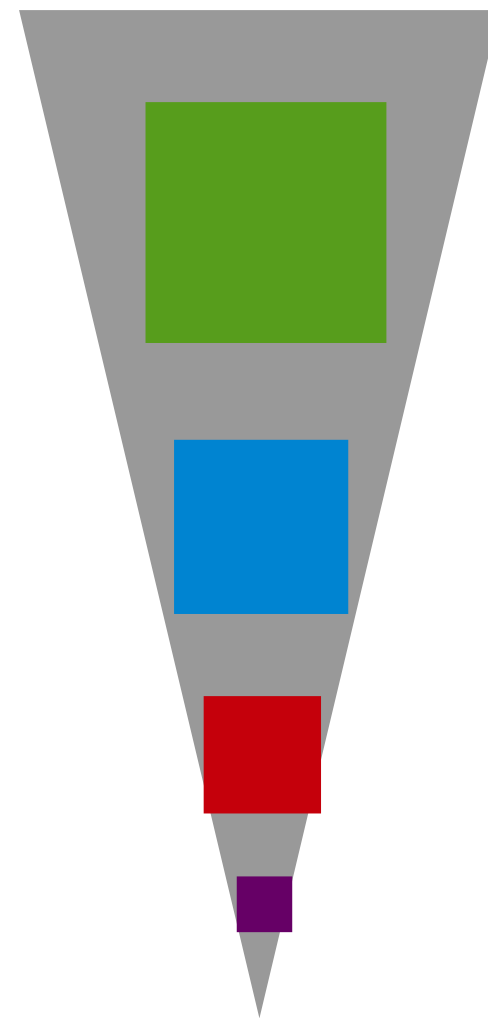
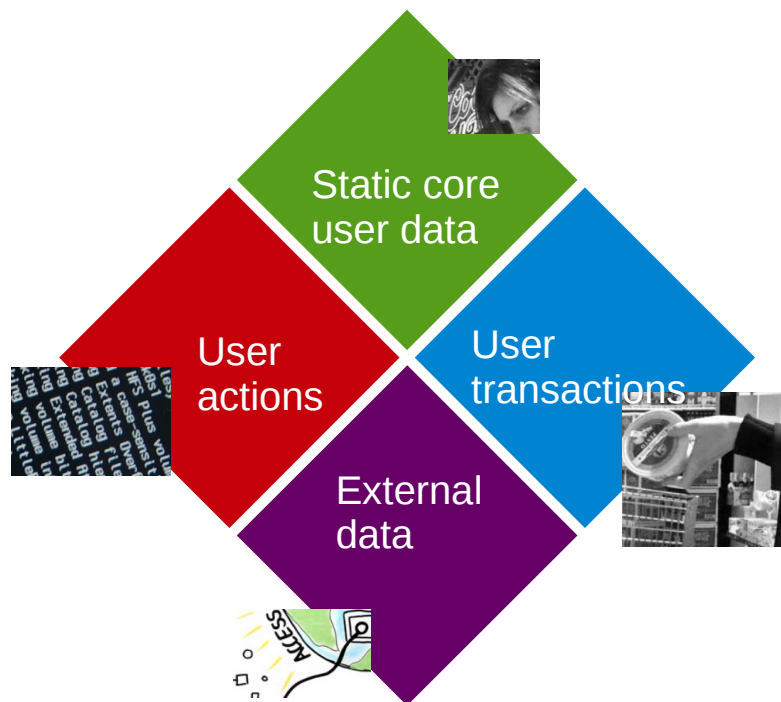
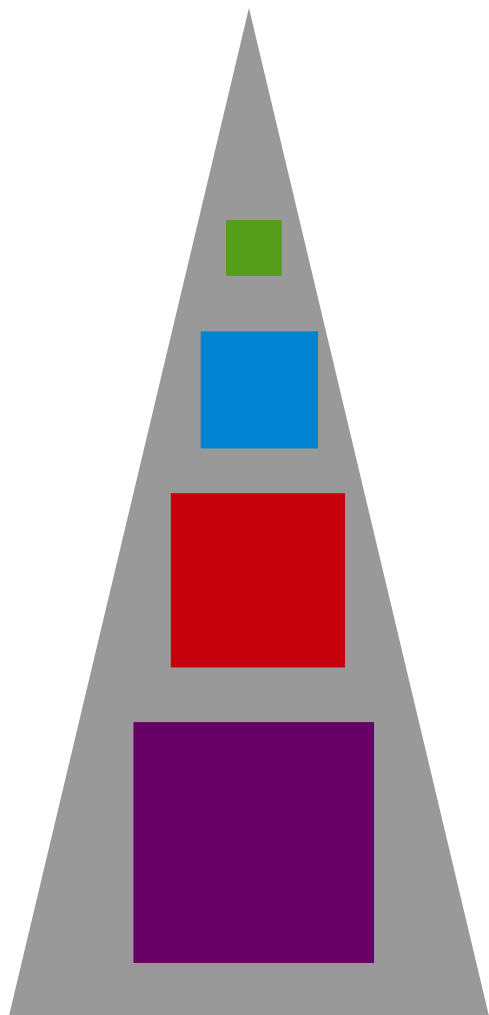
We are like it!!!

FRANKS WOZ ere os

I ♥ VIRGINIA



Data volume



Data quality

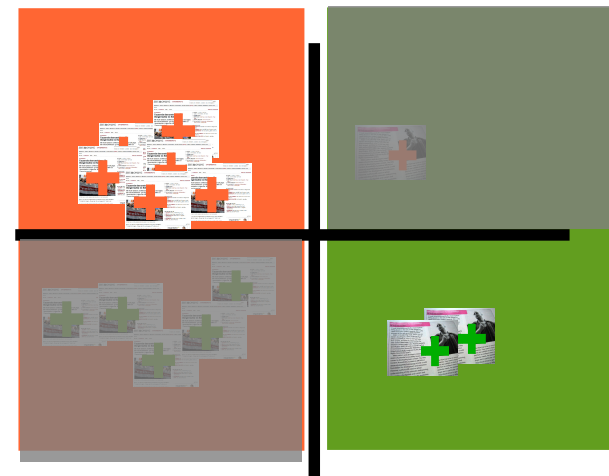


Image by wwarby: <http://www.flickr.com/photos/wwarby/3297205226>

Accuracy

$$ACC = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{false positive} + \text{false negative} + \text{true negative}}$$

- Problems:
 - What if class distribution is skewed?

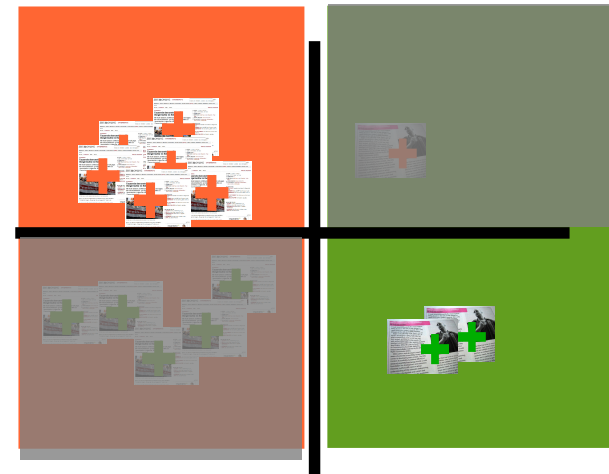


Precision/ Recall

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

- Problem:
 - Depends on decision threshold.



ROC Curves

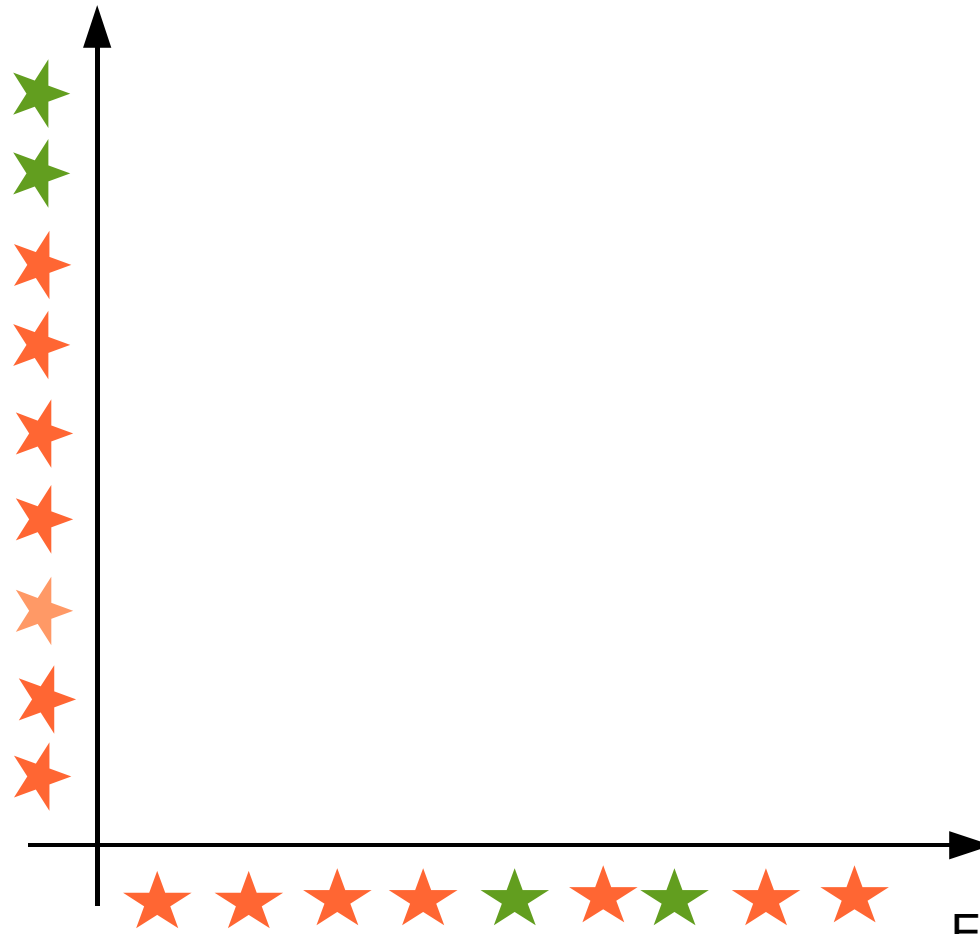


ROC Curves



ROC Curves

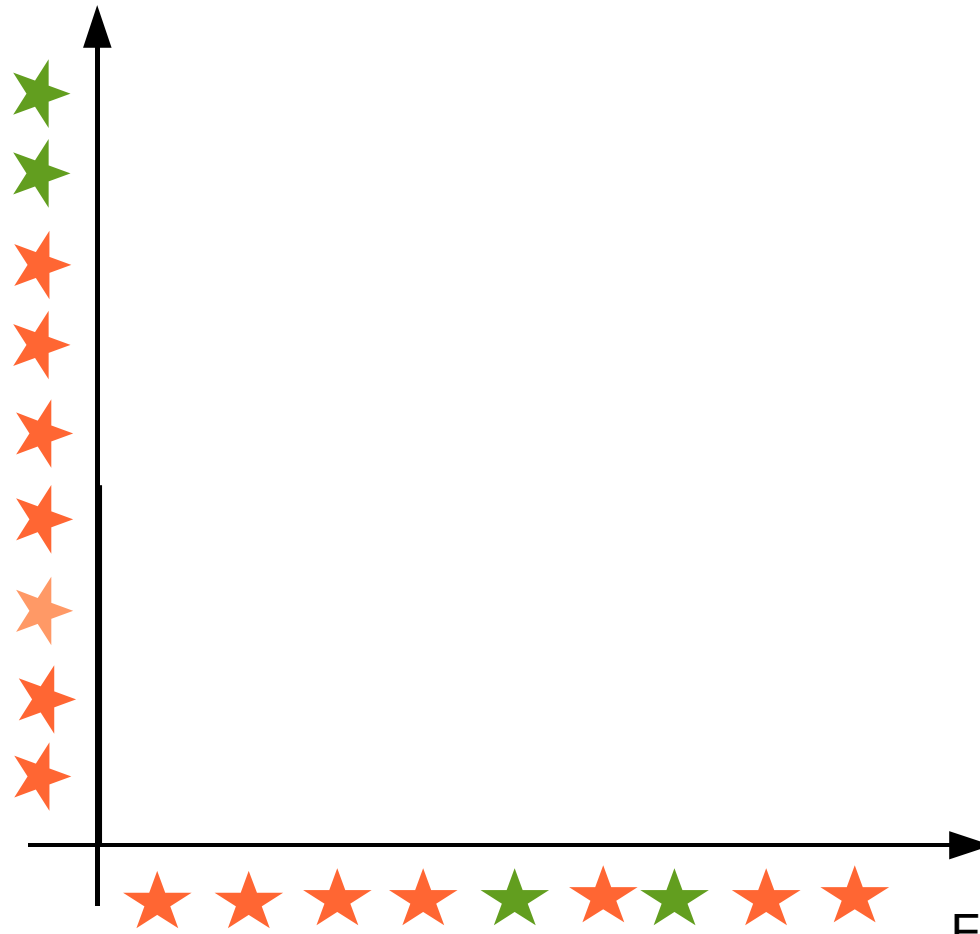
True orange rate



False orange rate

ROC Curves

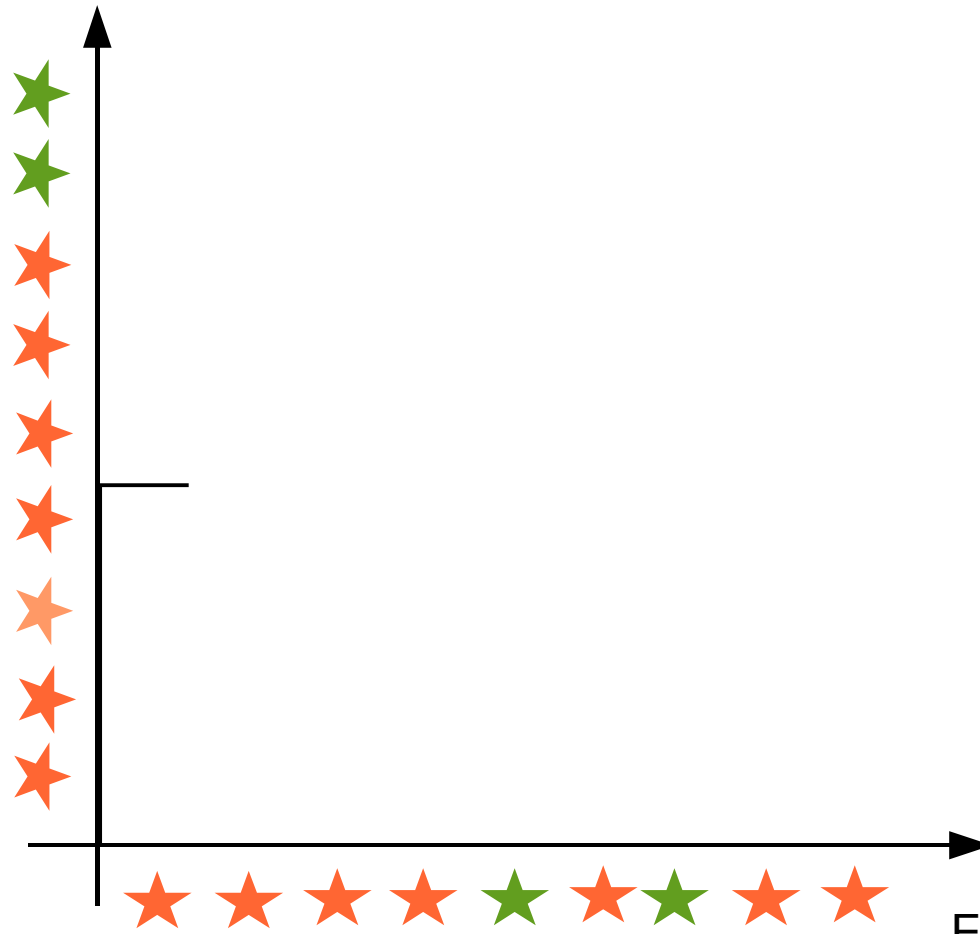
True orange rate



False orange rate

ROC Curves

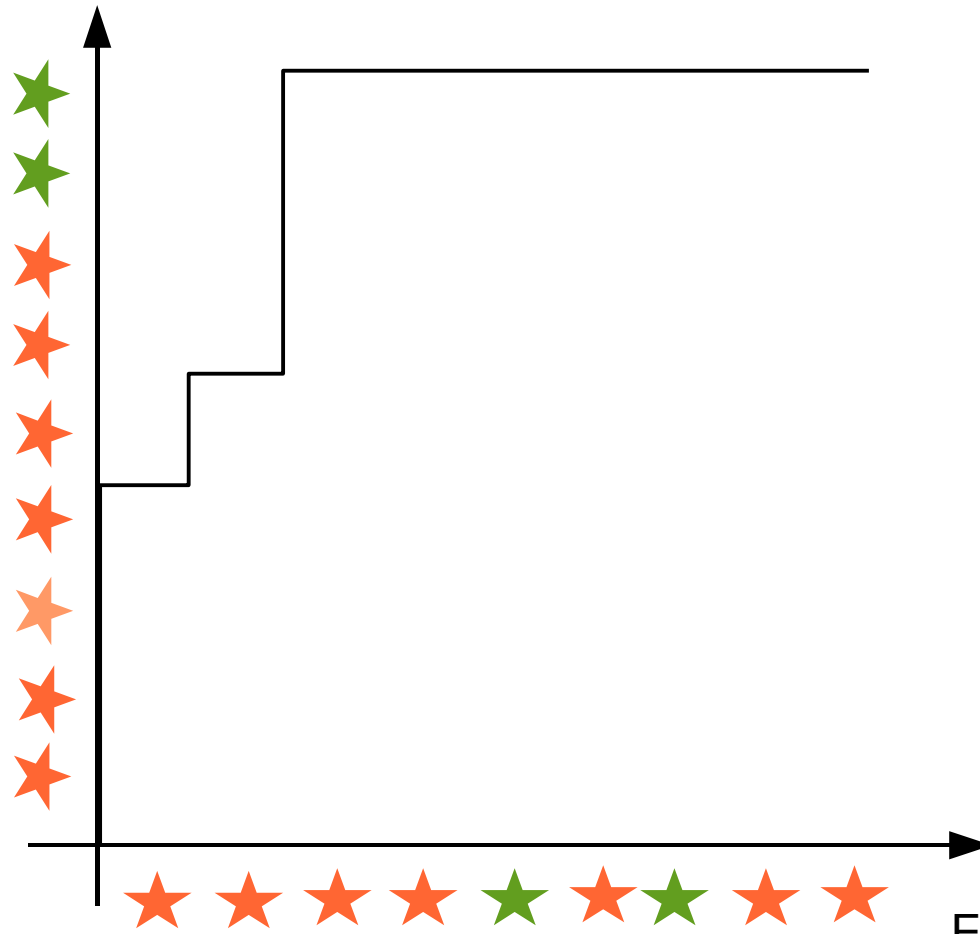
True orange rate



False orange rate

ROC Curves

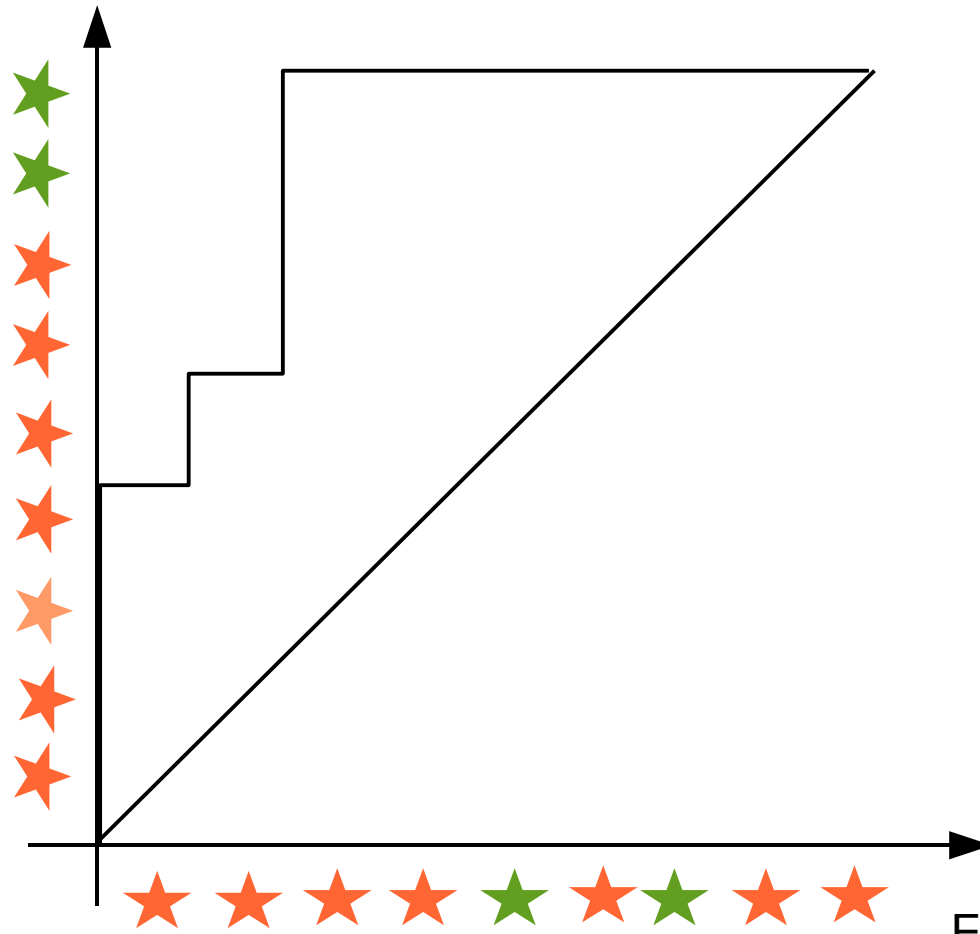
True orange rate



False orange rate

ROC Curves

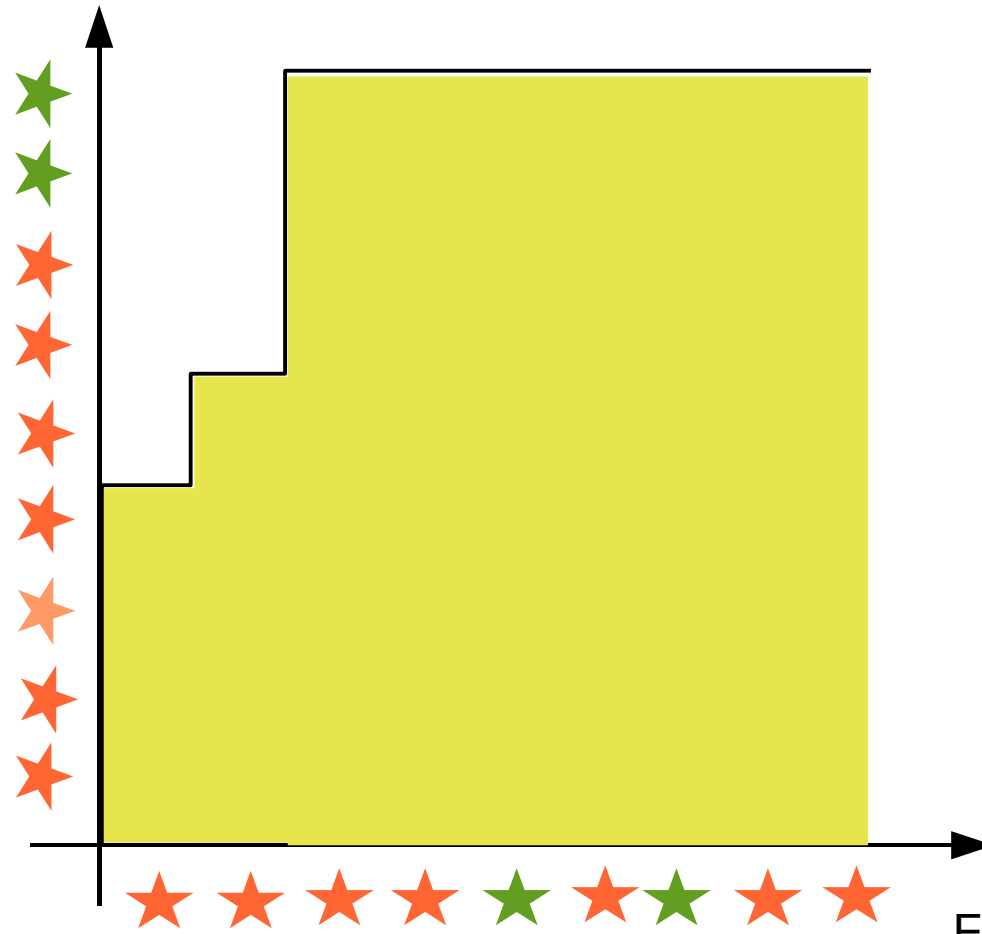
True orange rate



False orange rate

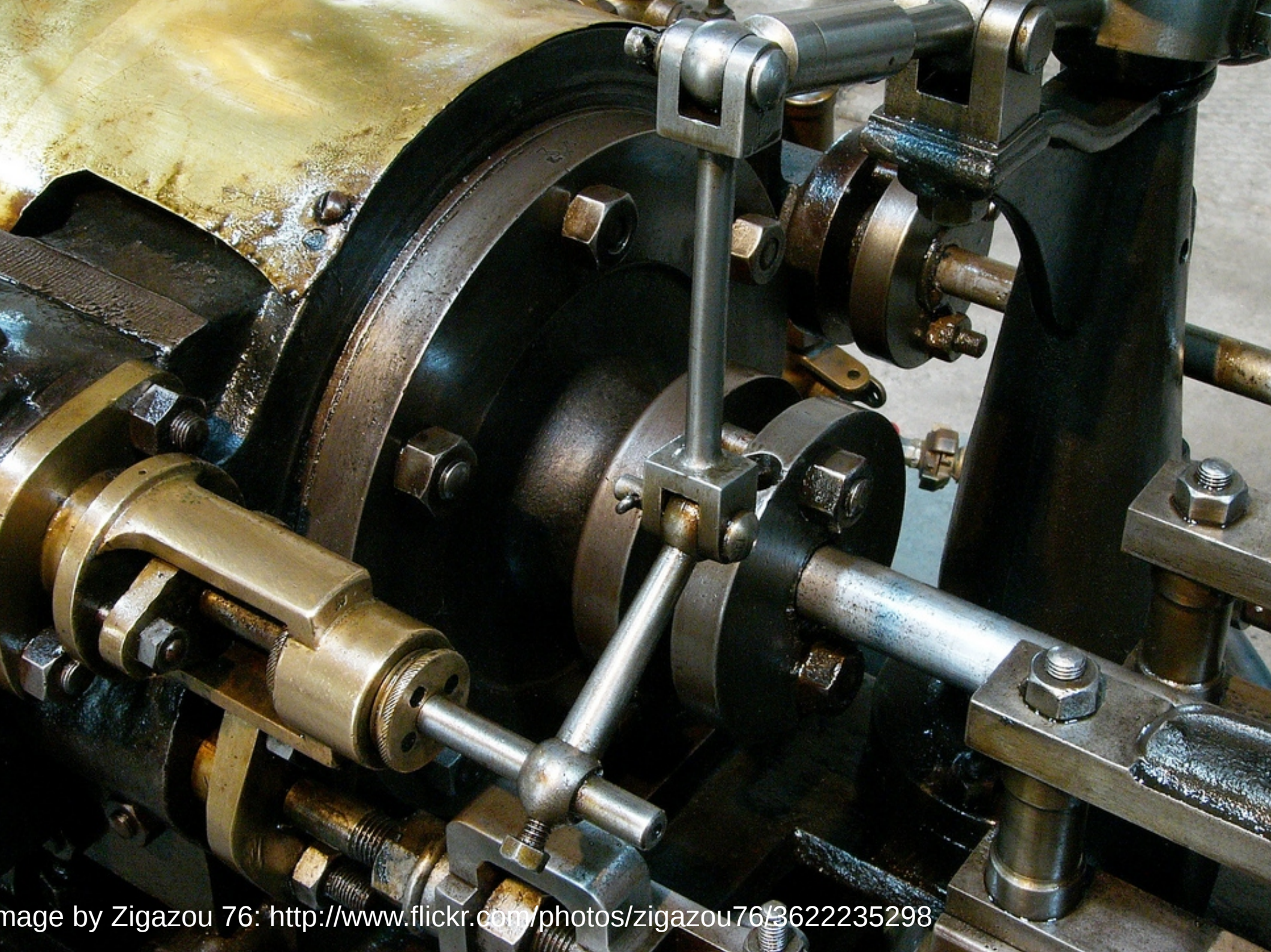
AUC – area under ROC

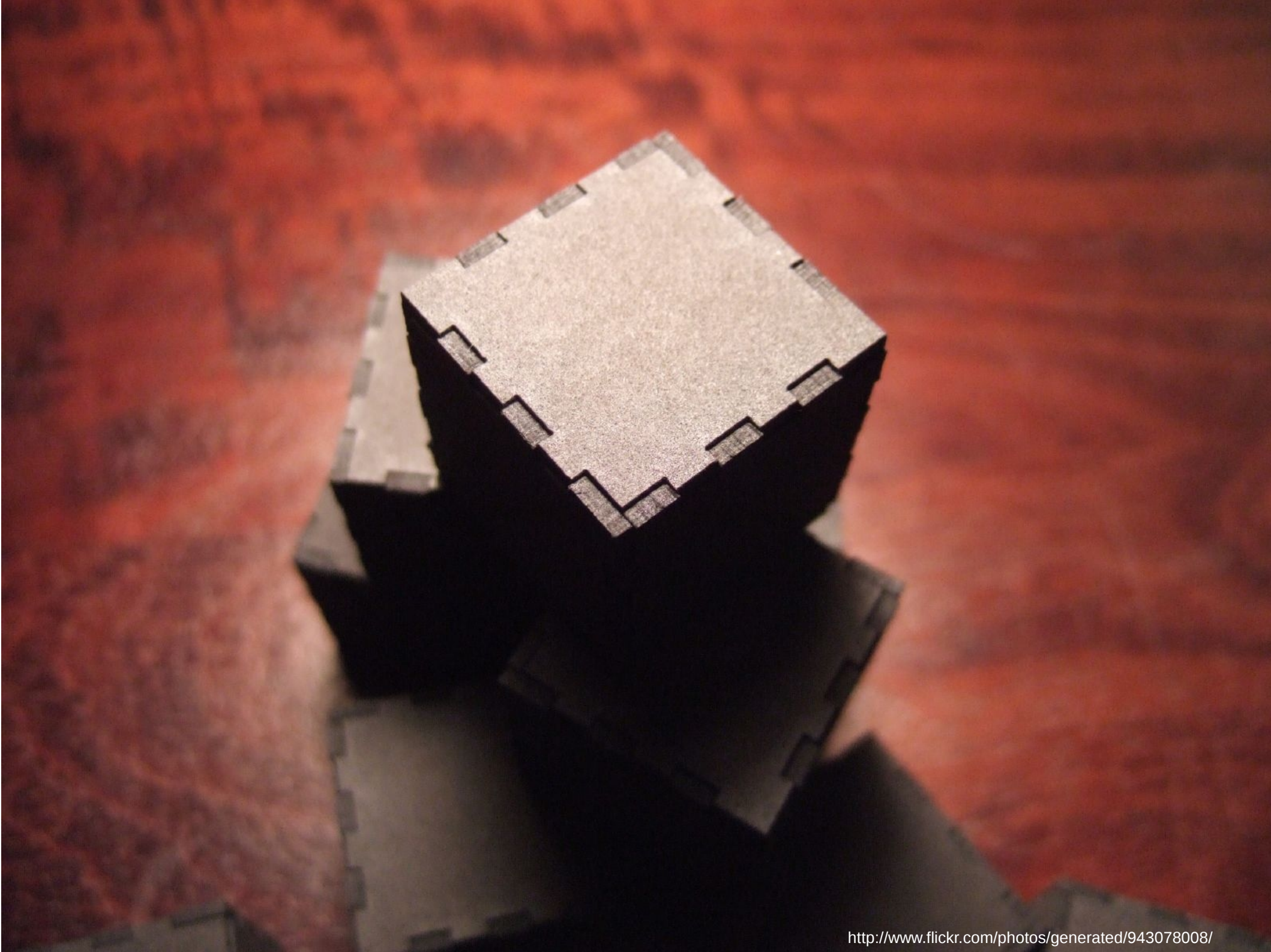
True orange rate



False orange rate

You want to check quality against real usage.







Requirements to get started

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

MacBook





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

Requirements to get started

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

MacBook



Image by Justin_a_glass: http://www.flickr.com/photos/justin_glass/3793492335

A photograph of a wooden pier extending into a calm blue lake. The pier is made of weathered wooden planks and has two vertical wooden posts on either side, connected by a horizontal beam. The lake is still, reflecting the sky. In the background, there are rolling hills and a small cluster of houses on a hillside. The sky is a clear, pale blue with a few wispy clouds. The overall mood is peaceful and serene.

Jumpstart your project with proven code.

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

A photograph of two people in a dimly lit room. On the left, a person with dark hair is seen in profile, resting their head on their hand and gesturing with their other hand. On the right, a person with glasses and long hair is sitting in a chair, leaning forward with their chin resting on their hand, listening intently. The background is a plain wall with a window showing some light. The overall mood is contemplative and focused.

Discuss ideas and problems online.

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

WARNING





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.



Register with
geecon_12
to get a 10% discount.





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.



New requirements:

- Find deficiencies in online reviews
- Find preferred features online

Data based features:

- Link to external content.
- Show what I'm interested in based on my social network.

Peter Norvig (paraphrase): Rather than argue about whether this algorithm is better than that algorithm, all you have to do is get ten times more training data. And now all of a sudden, the worst algorithm ... is performing better than the best algorithm on less training data.

Worry about the data first before you worry about the algorithm.



Google code **protobuf**
Protocol Buffers - Google's data interchange format



Java API
Pipes C++ API
Streaming
Dumbo for Python
...



Cascading



Export much smaller
result set to database
and use standard tools.

Use custom visualization.

Export to file format
compatible with your
favourite graphing tool.