TM

Apache Mahout in context

Choosing the right tool for your data analysis task

Isabel Drost-Fromm

Software Engineer at Nokia Maps

Member of the Apache Software Foundation

Co-Founder of Berlin Buzzwords and
Berlin Apache Hadoop GetTogether

Co-founder of Apache Mahout

https://cwiki.apache.org/confluence/display/MAHOUT/Powered+By+Mahout

… provide your own success story online.

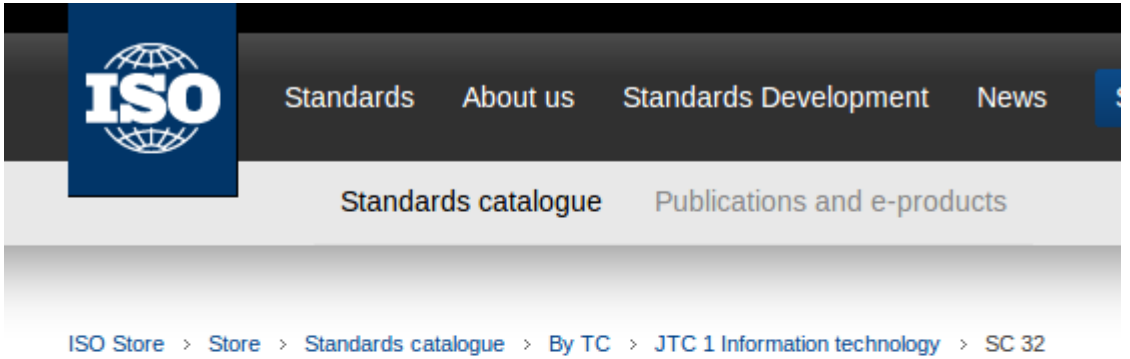
… with input from Sebastian Schelter, Steffen Bickel, Zeno Gantner, Stefan Pohl, Shannon Quinn, and others

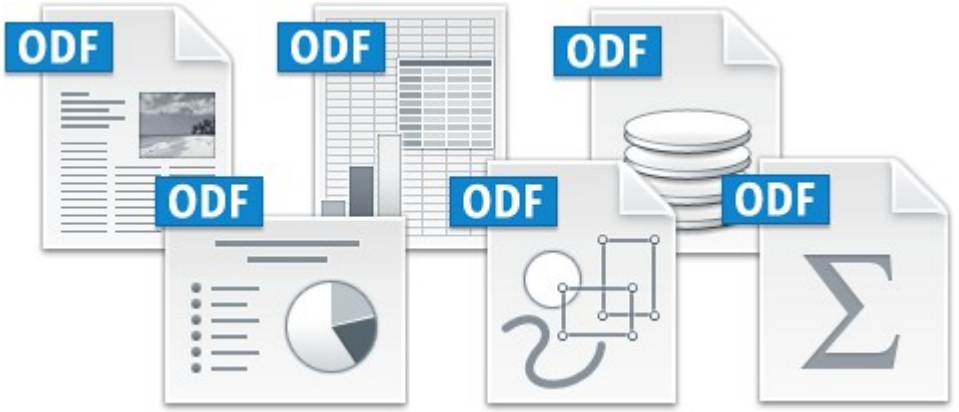

rapid-i
REPORT THE FUTURE


MATLAB
MAJOR UPDATE
The Language of Technical Computing


SciPy.org

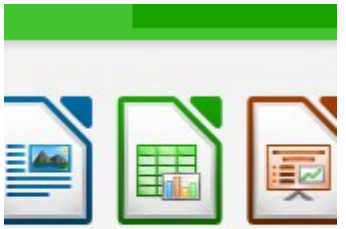

tableau
SOFTWARE


Standards
About us
Standards Development
News
Standards catalogue
Publications and e-products
ISO Store › Store › Standards catalogue › By TC › JTC 1 Information technology › SC 32
Standards catalogue
Browse by ICS
Browse by TC
JTC 1/SC 32 - Data management and interchange



Excel 2010

Buy with Office | Download a trial



ibreOffice is the power-pa

TM


mahout

Data science is a discipline that incorporates varying elements and **builds on techniques and theories from many fields**, including Math, Statistics, Data Engineering, Pattern Recognition and Learning, Advanced Computing, Visualization, Uncertainty Modeling, Data Warehousing, and high performance computing **with the goal of extracting meaning from data and creating data products**.

Data Science seeks to use all available and relevant data to effectively **tell a story** that can be easily understood by non-practitioners.


Image taken at Berlin Buzzwords by photomic

Trends

Web Search Interest: **servlet, django, node.js**. Worldwide, 2004 - present.

Explore trends

Hot searches

Search terms

- servlet
- django
- node.js

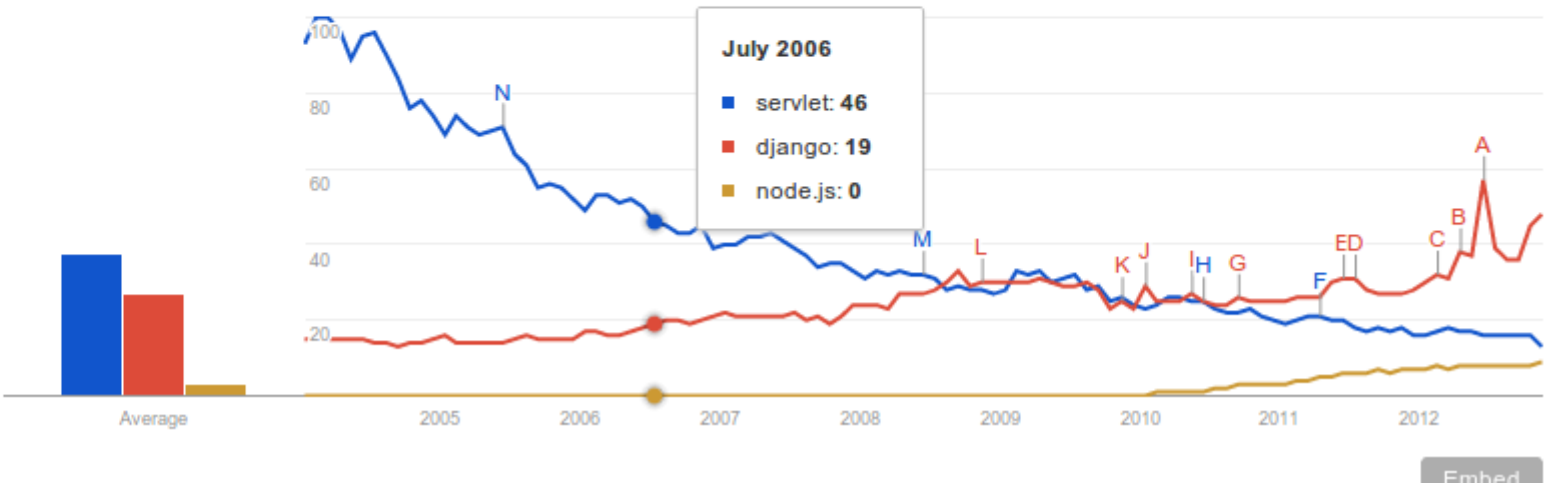
+ Add term

Other comparisons

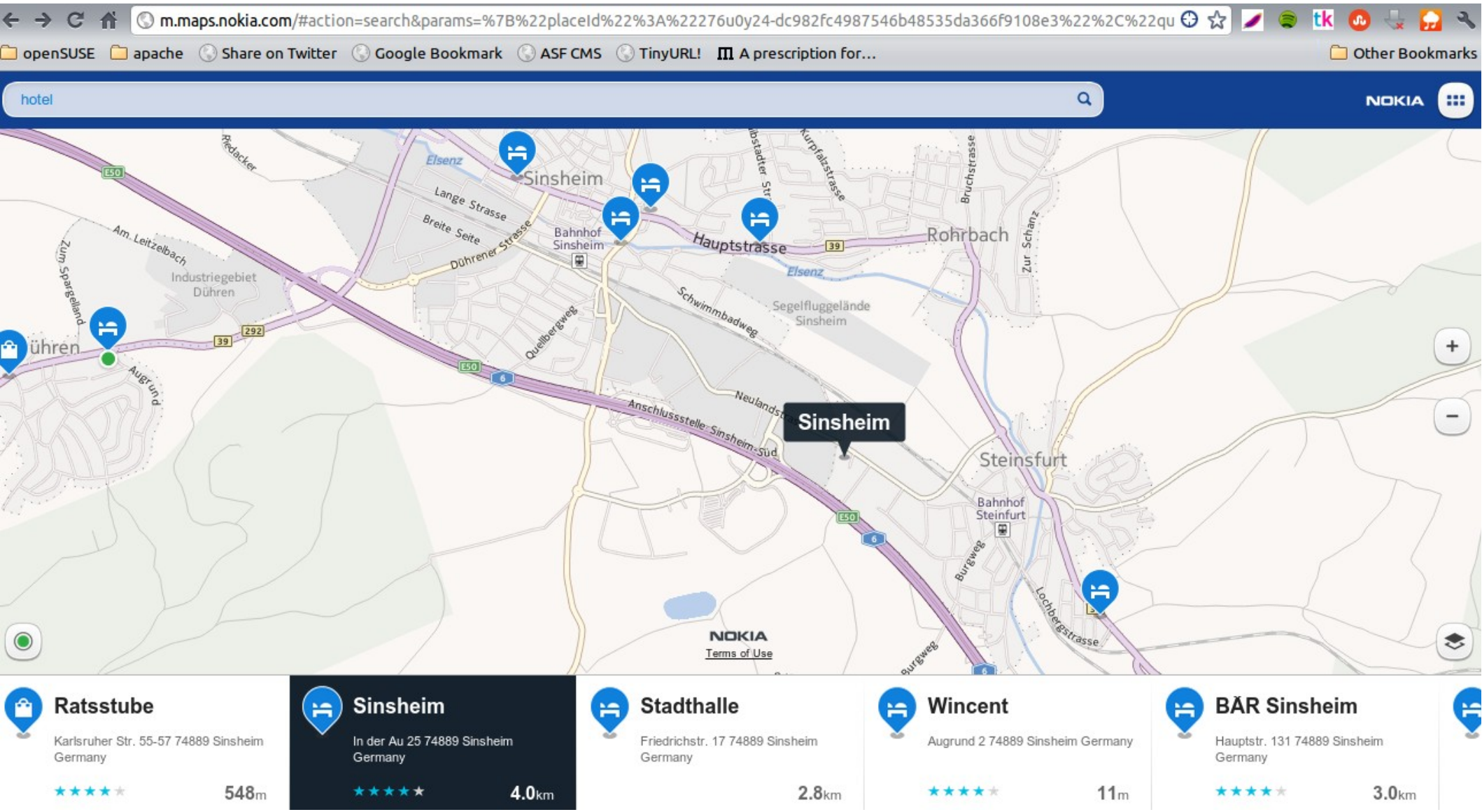
Interest over time

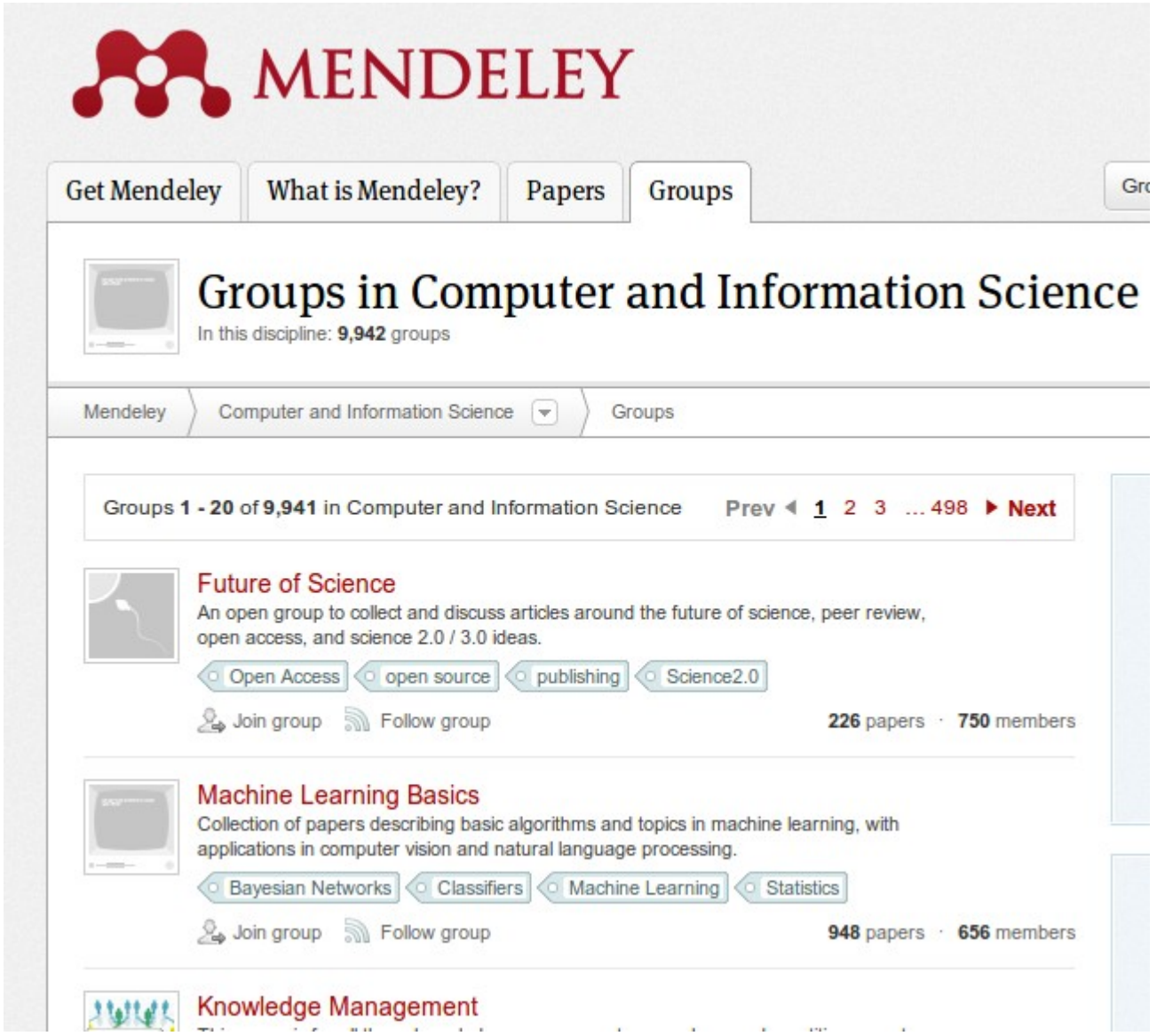
The number 100 represents the peak search volume

News headlines Forecast



Embed


m.maps.nokia.com/#action=search&params=%7B%22placeId%22%3A%22276u0y24-dc982fc4987546b48535da366f9108e3%22%2C%22qu
openSUSE
apache
Share on Twitter
Google Bookmark
ASF CMS
TinyURL!
A prescription for...
Other Bookmarks
hotel
NOKIA
Sinsheim
Elsenz
Lange Strasse
Breite Seite
Dühren
Bahnhof Sinsheim
Hauptstrasse
Rohrbach
Bruchstrasse
Kurpfalzstrasse
Zur Schanz
Am. Leitzelbach
Industriegebiet Dühren
Zum Spargelland
Augrund
Schwimmbadweg
Quellbergweg
Segelfluggelände Sinsheim
Neulandstr.
Anschlussstelle Sinsheim-Süd
Steinsfurt
Bahnhof Steinfurt
Burgweg
Lochbergstrasse
NOKIA
Terms of Use
Ratsstube
Karlsruher Str. 55-57 74889 Sinsheim Germany
548m
Sinsheim
In der Au 25 74889 Sinsheim Germany
4.0km
Stadthalle
Friedrichstr. 17 74889 Sinsheim Germany
2.8km
Wincent
Augrund 2 74889 Sinsheim Germany
11m
BÄR Sinsheim
Hauptstr. 131 74889 Sinsheim Germany
3.0km


MENDELEY
Get Mendeley
What is Mendeley?
Papers
Groups
Groups in Computer and Information Science
In this discipline: 9,942 groups
Mendeley
Computer and Information Science
Groups
Groups 1 - 20 of 9,941 in Computer and Information Science
Prev 1 2 3 ... 498 Next
Future of Science
An open group to collect and discuss articles around the future of science, peer review, open access, and science 2.0 / 3.0 ideas.
Open Access
open source
publishing
Science2.0
Join group
Follow group
226 papers · 750 members
Machine Learning Basics
Collection of papers describing basic algorithms and topics in machine learning, with applications in computer vision and natural language processing.
Bayesian Networks
Classifiers
Machine Learning
Statistics
Join group
Follow group
948 papers · 656 members
Knowledge Management

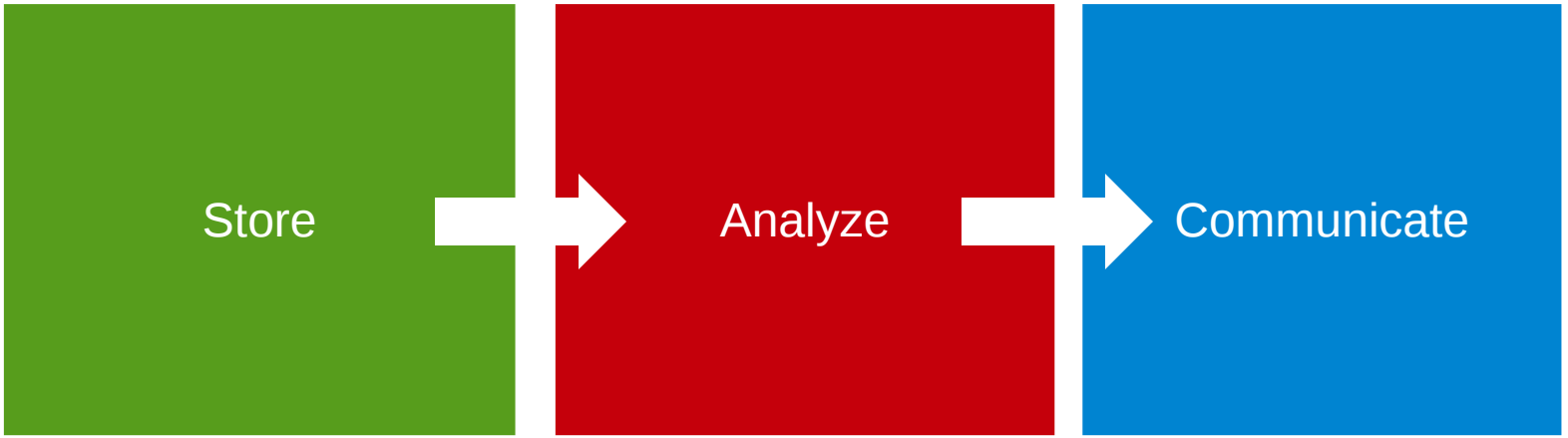

Store
Analyze
Communicate

Image by jasondevilla
http://www.flickr.com/photos/jasondv/91960897/

How a linear classifier sees data

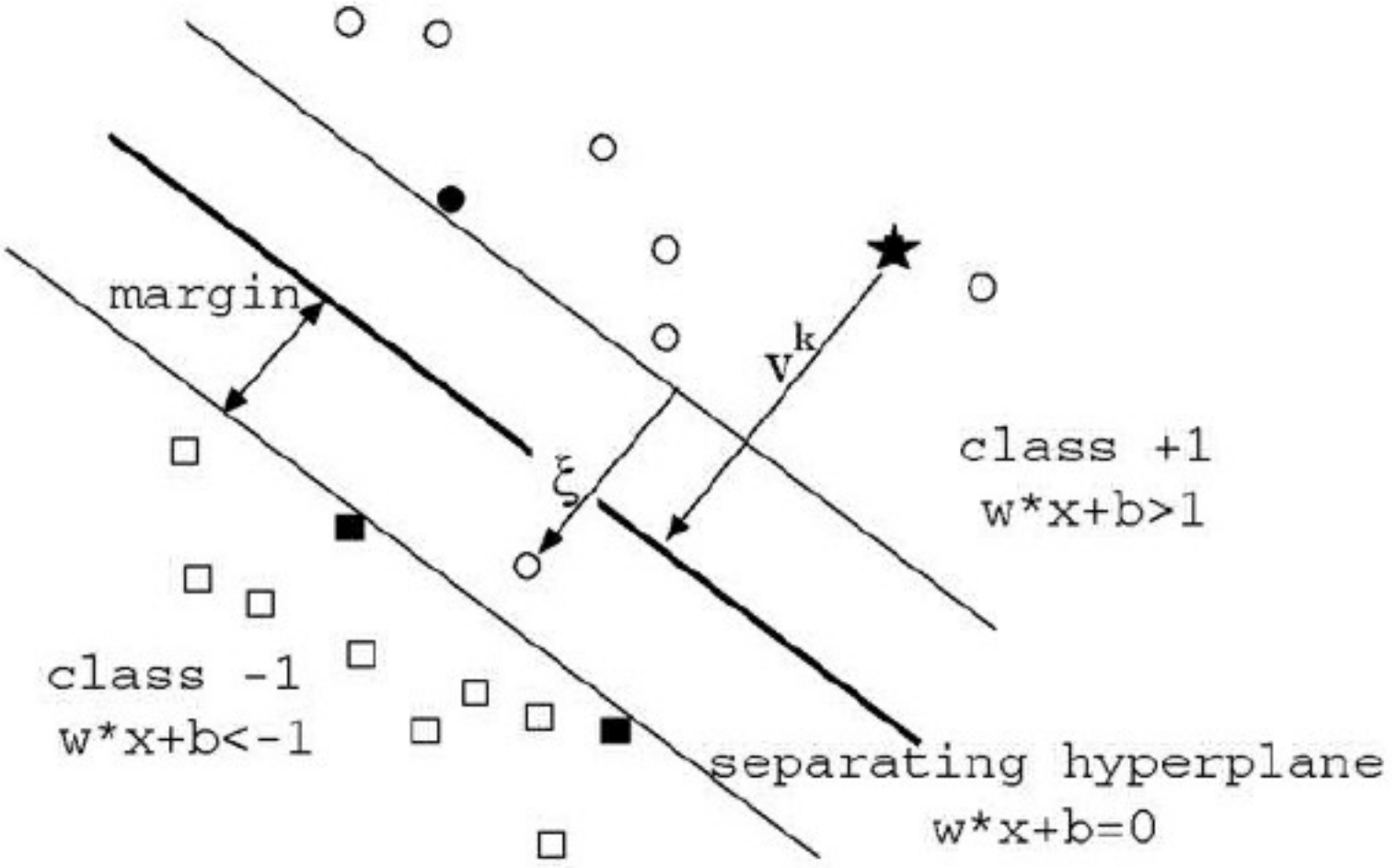



Image by ZapTheDingbat (Light meter)
http://www.flickr.com/photos/zapthedingbat/3028168415
DIN
LENINGRAD

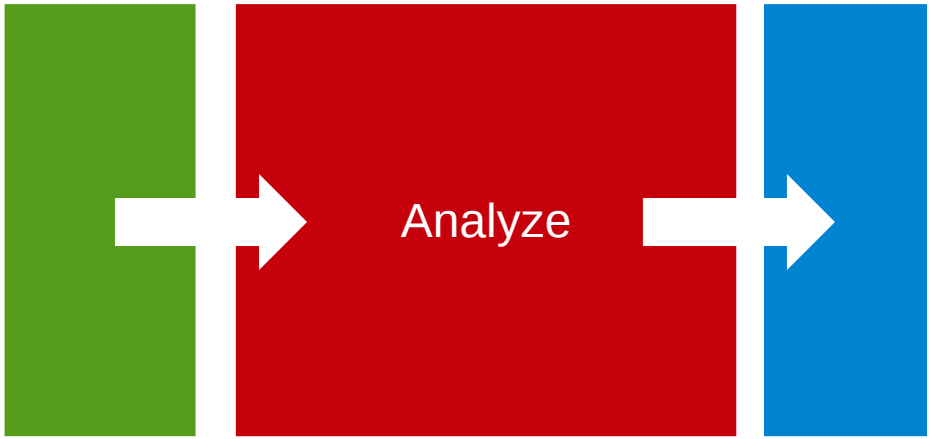
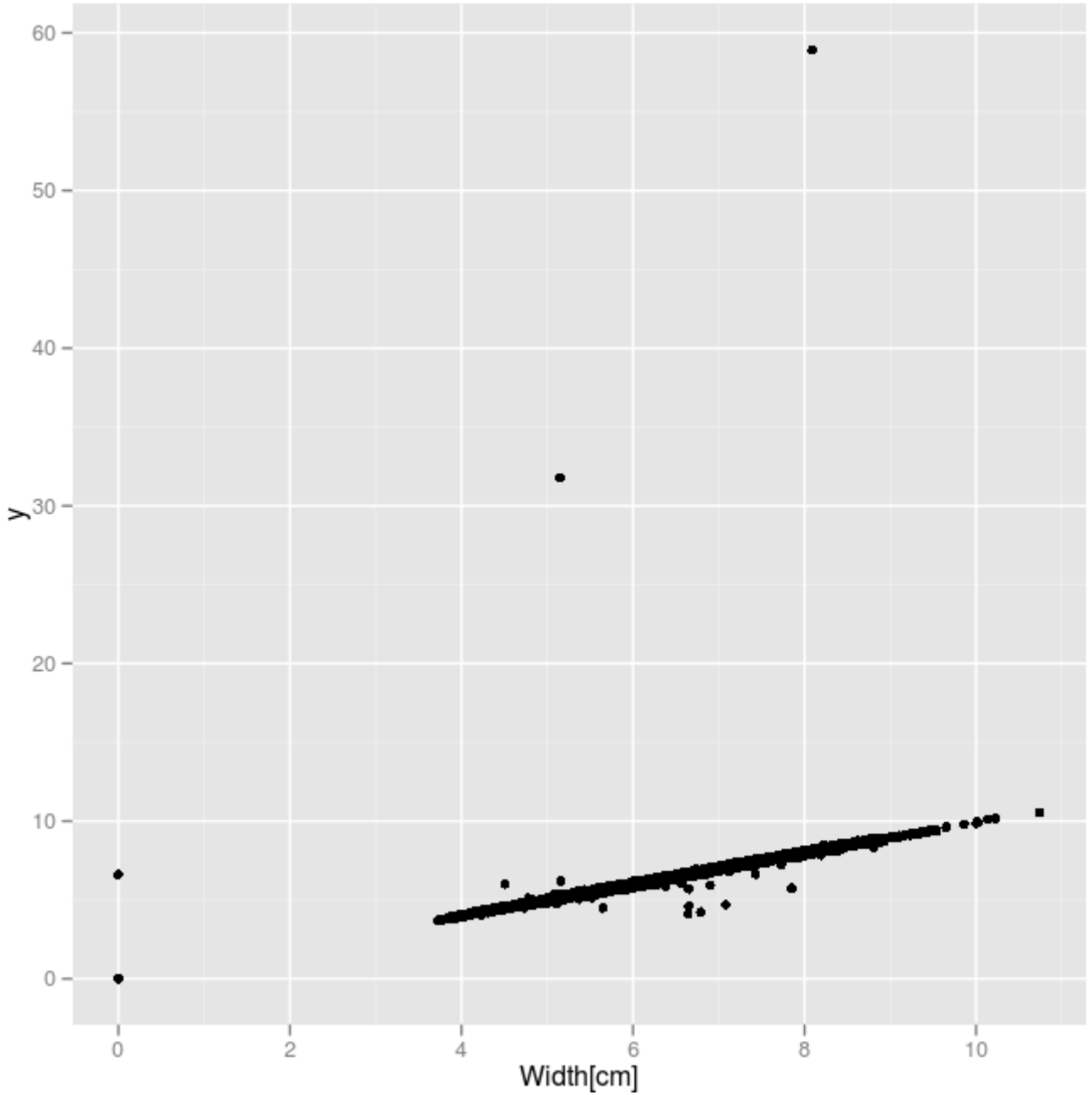

Analyze

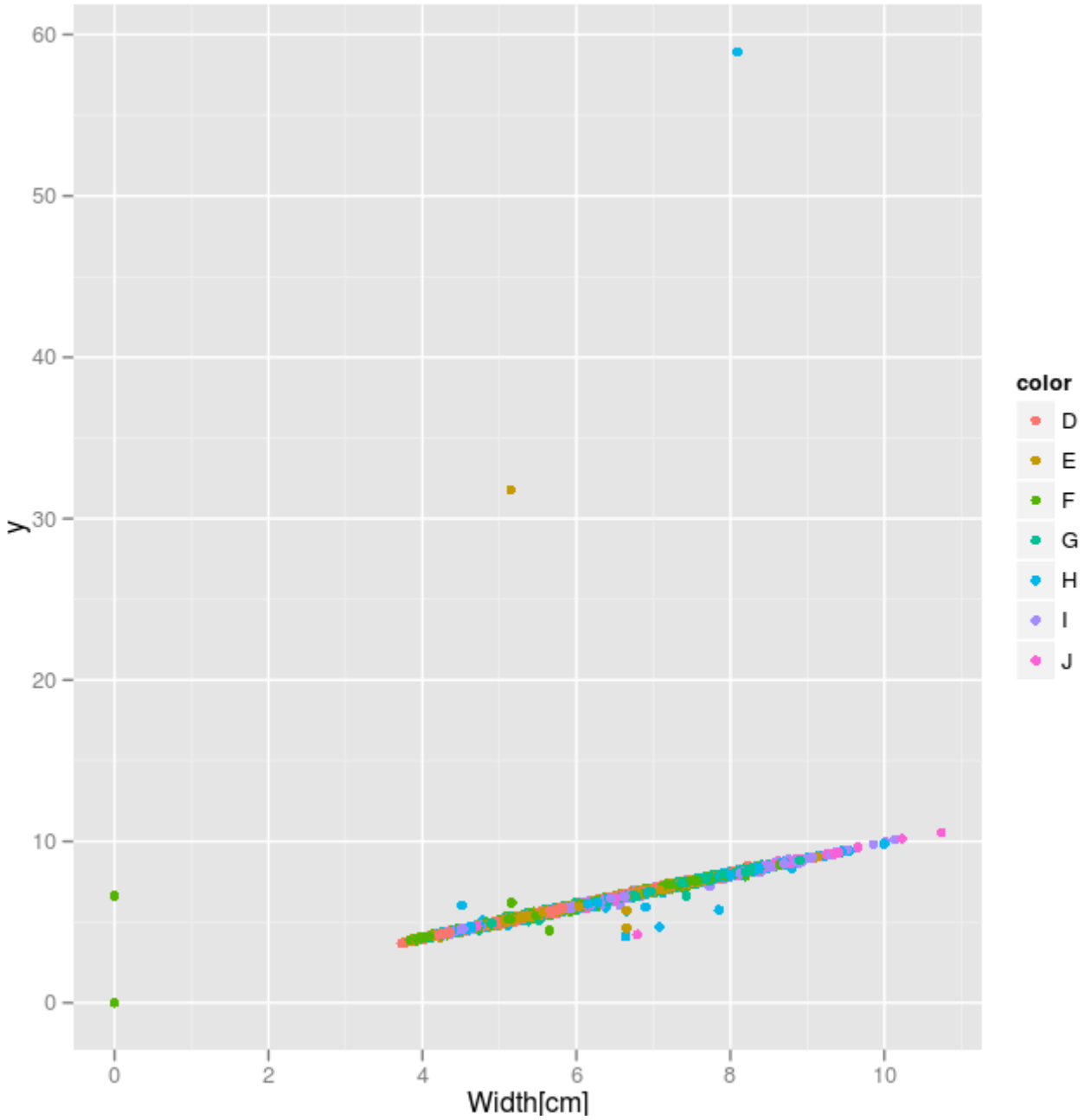

Image by Thilo Fromm
https://plus.google.com/photos/102813492714620417611/albums/5781077220806954385/5781079983265400098
Taken in Yosemite Park 2011.

```
> head(diamonds)
  carat       cut color clarity depth table price    x    y    z
1  0.23     Ideal     E     SI2  61.5    55   326 3.95 3.98 2.43
2  0.21   Premium     E     SI1  59.8    61   326 3.89 3.84 2.31
3  0.23      Good     E     VS1  56.9    65   327 4.05 4.07 2.31
4  0.29   Premium     I     VS2  62.4    58   334 4.20 4.23 2.63
5  0.31      Good     J     SI2  63.3    58   335 4.34 4.35 2.75
6  0.24 Very Good     J    VVS2  62.8    57   336 3.94 3.96 2.48
```

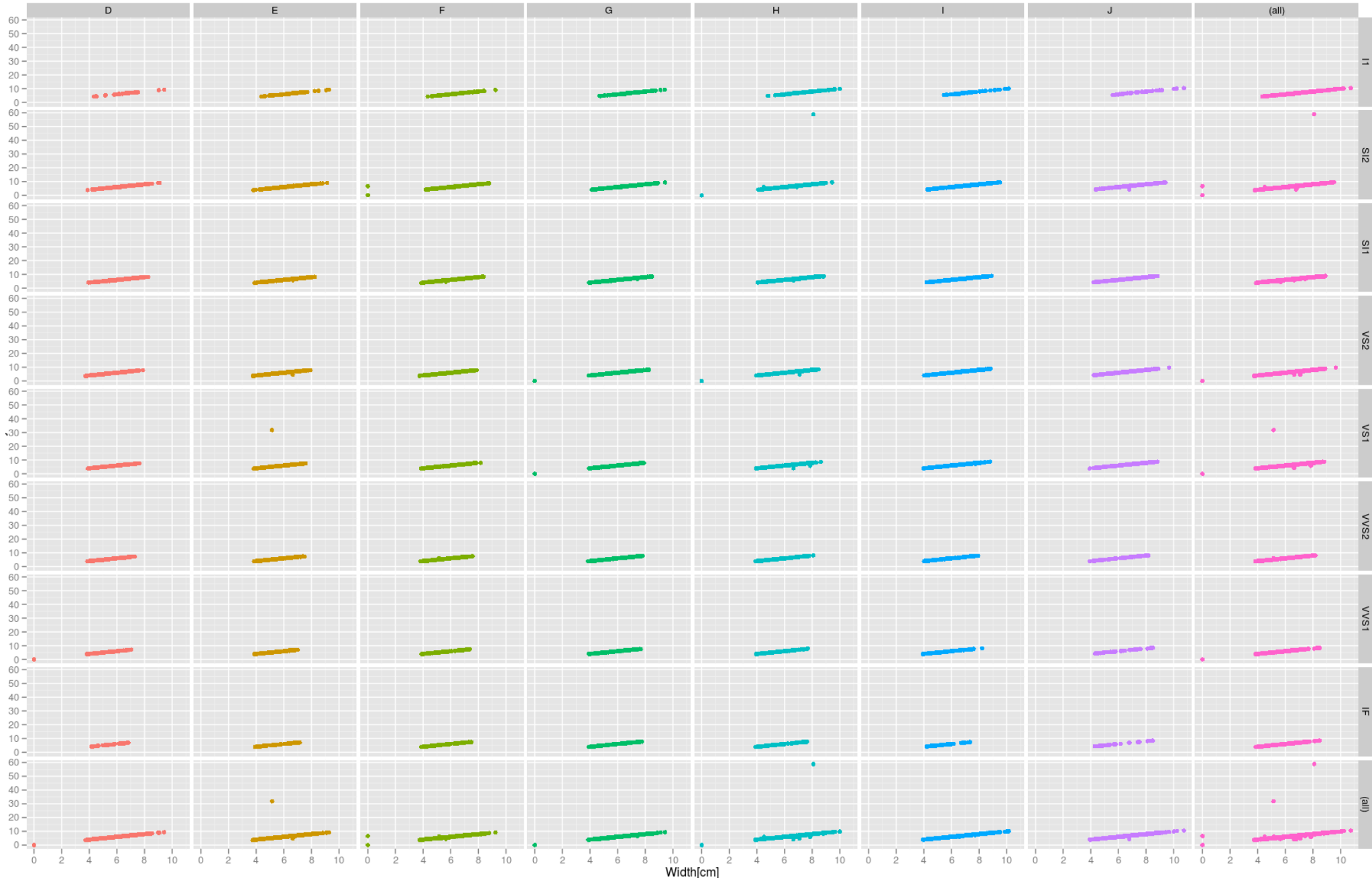
```
> summary(diamonds)
     carat               cut        color        clarity          depth      
 Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00  
 1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00  
 Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80  
 Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75  
 3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50  
 Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00  
                                    J: 2808   (Other): 2531                  
```



Ggplot2 examples courtesy of Stefan Pohl.



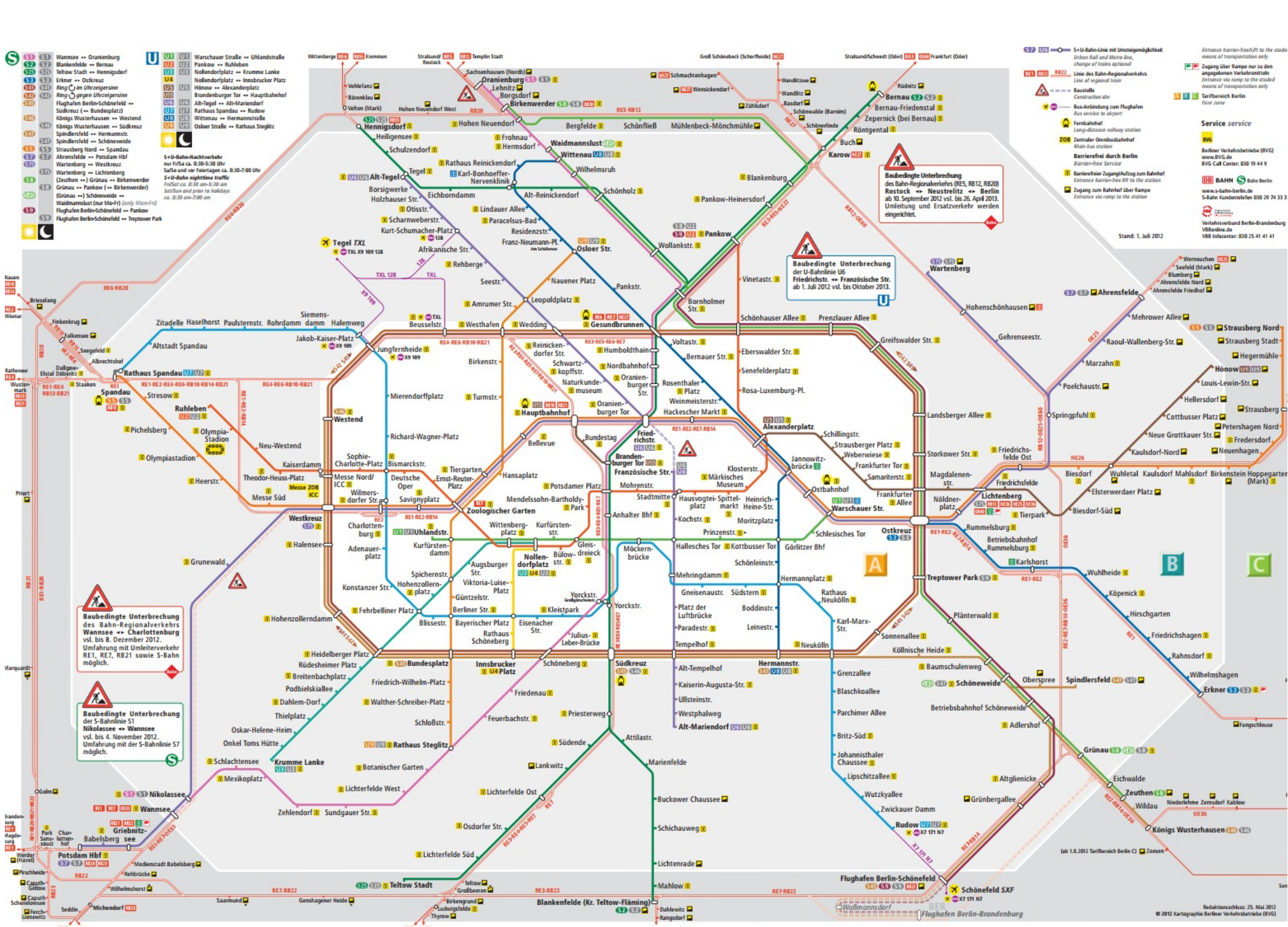
Ggplot2 examples courtesy of Stefan Pohl.



Ggplot2 examples courtesy of Stefan Pohl.

```
> R

> library(ggplot2)

> head(diamonds)

> summary(diamonds)


> ggplot(diamonds, aes(x,y)) + geom_point() + xlab('Width[cm]')


> ggplot(diamonds, aes(x,y, color=color)) + geom_point() +
xlab('Width[cm]')


> ggplot(diamonds, aes(x,y, color=color)) + geom_point() +
xlab('Width[cm]') +facet_grid(clarity ~ color, margins=T)
```

Ggplot2 examples courtesy of Stefan Pohl.



Map of Berlin public transport lines.

Image taken by Thilo Fromm.
https://plus.google.com/photos/102813492714620417611/albums/5781077220806954385/5781079910600488146
Kommunikationsmuseum Berlin.

Instance*

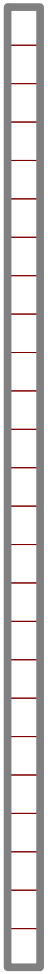
(sometimes also called example, item, or in databases a row)

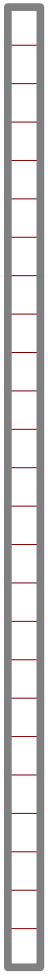
Feature*

(sometimes also called attribute, signal, predictor, co-variate, or column in databases)

Label*

(sometimes also called class, target variable)





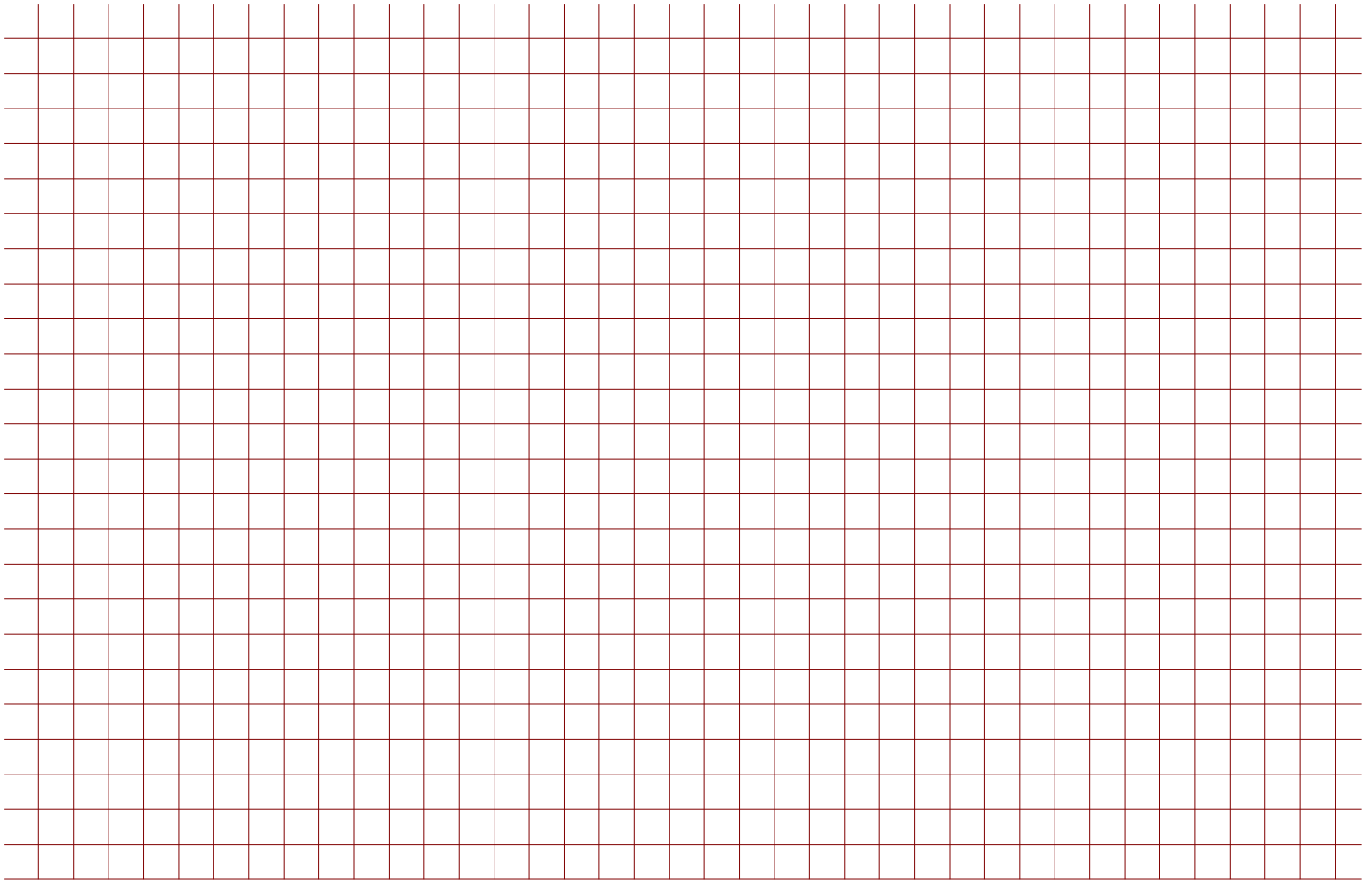
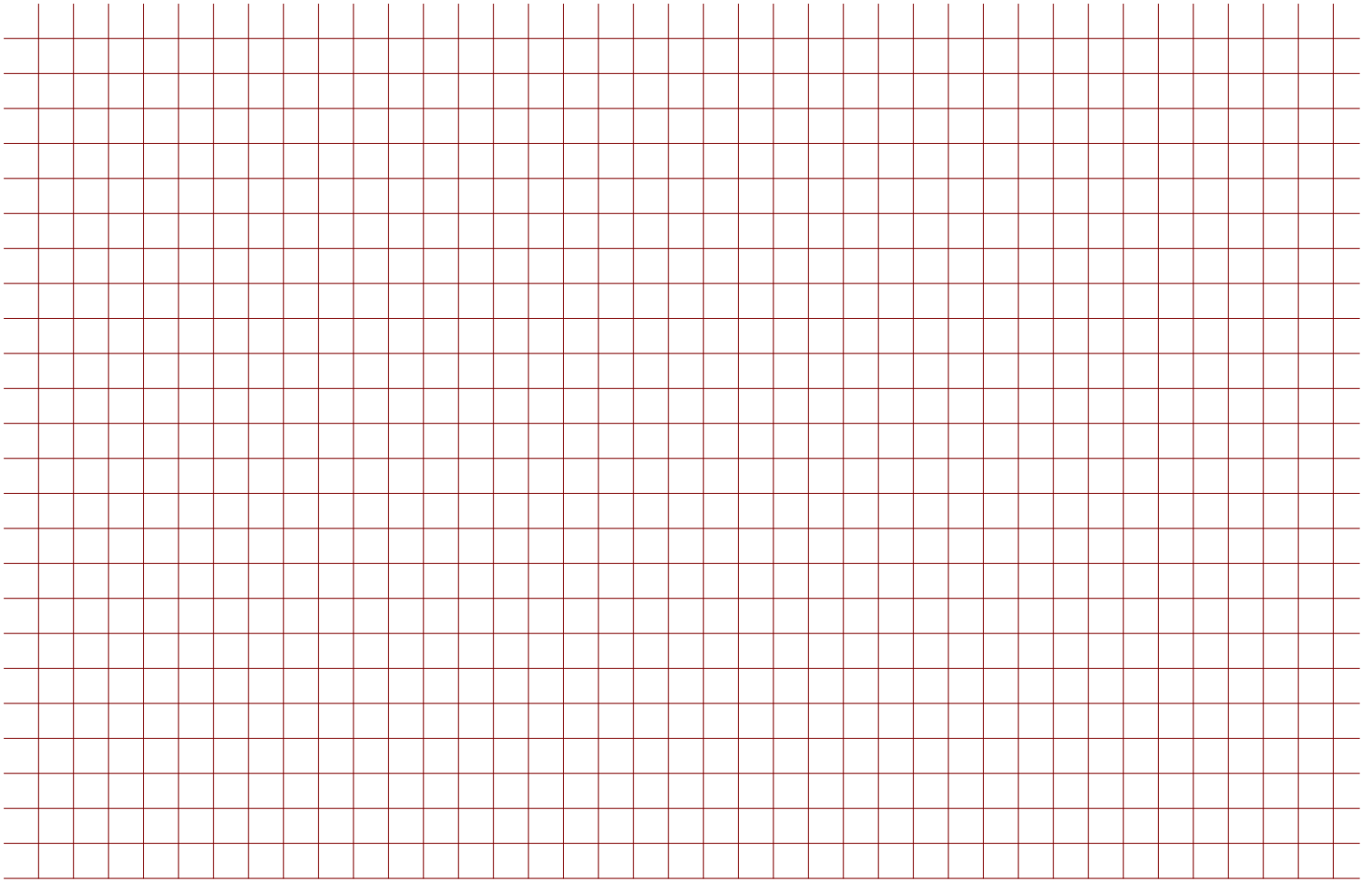
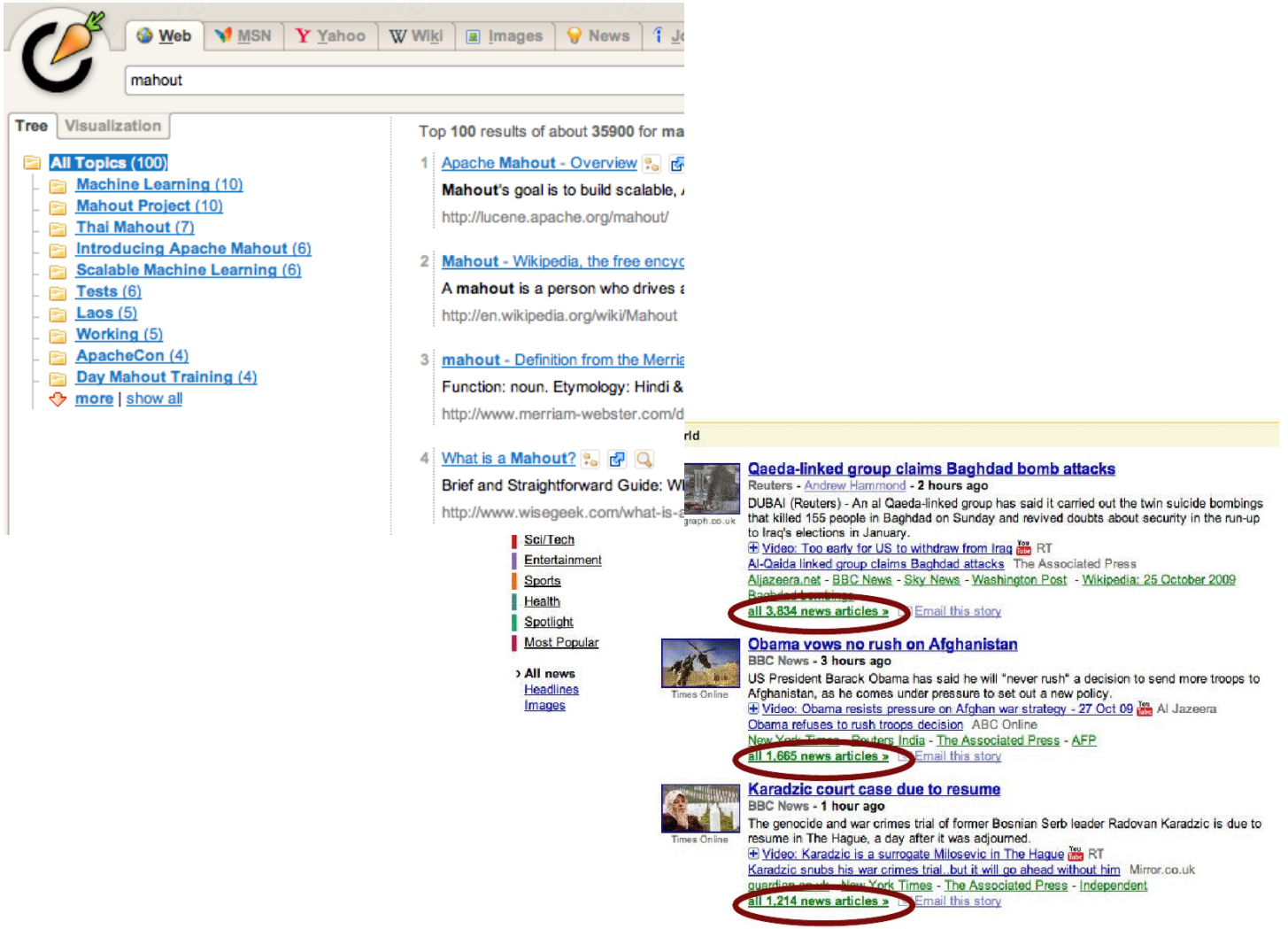
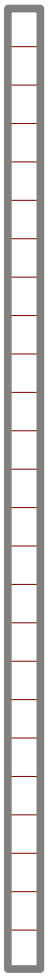
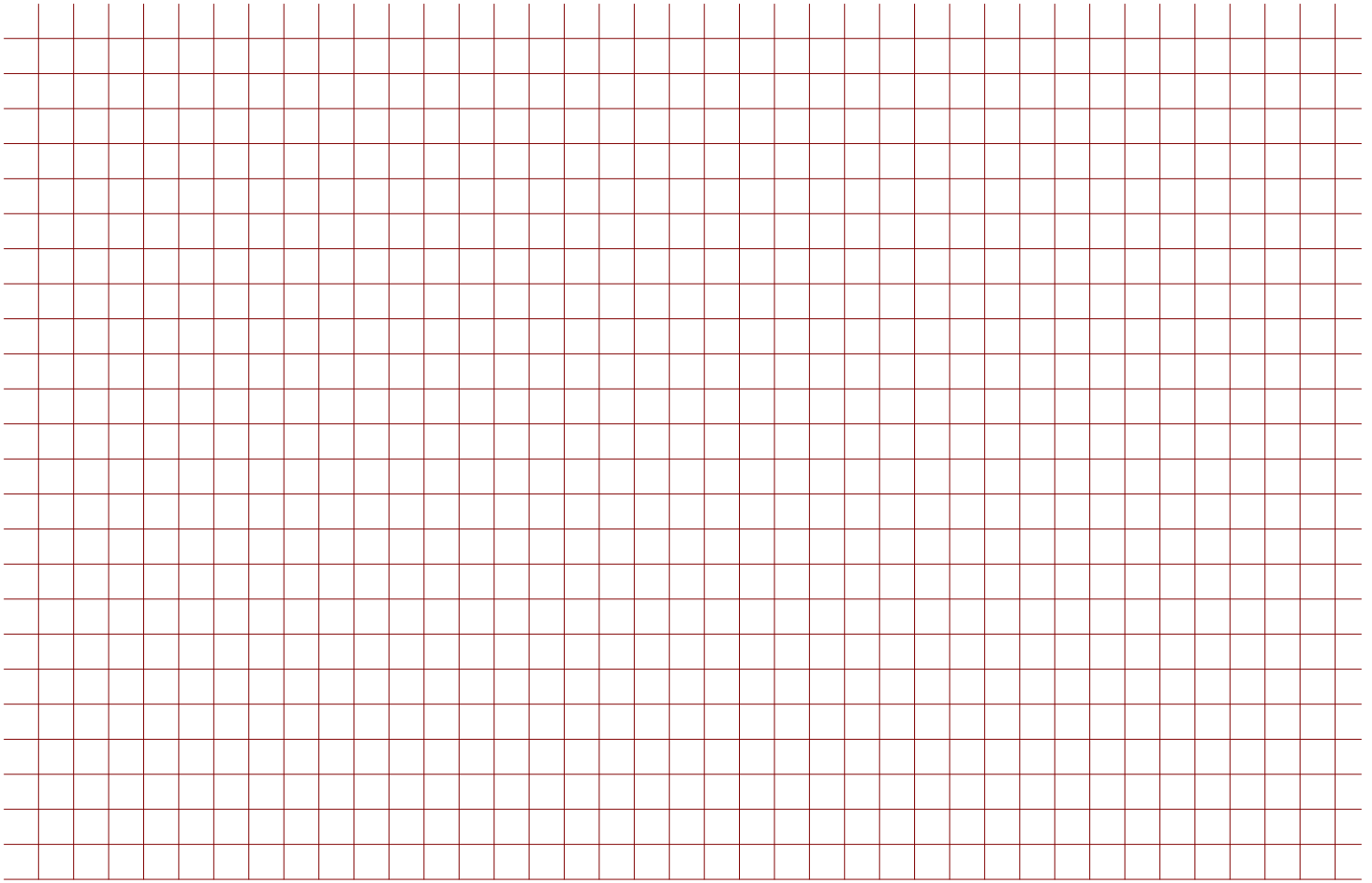



Image taken in Lisbon/ Portugal.

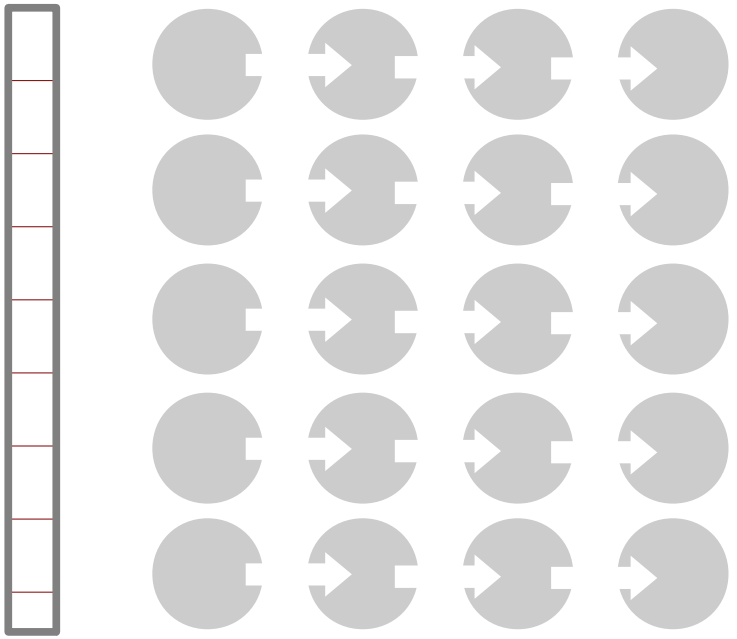



Web
MSN
Yahoo
Wiki
Images
News
mahout
Tree
Visualization
All Topics (100)
Machine Learning (10)
Mahout Project (10)
Thai Mahout (7)
Introducing Apache Mahout (6)
Scalable Machine Learning (6)
Tests (6)
Laos (5)
Working (5)
ApacheCon (4)
Day Mahout Training (4)
more | show all
Top 100 results of about 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 Merria
Function: noun. Etymology: Hindi &
http://www.merriam-webster.com/d
4 What is a Mahout?
Brief and Straightforward Guide: Wh
http://www.wisegeek.com/what-is-a
Sci/Tech
Entertainment
Sports
Health
Spotlight
Most Popular
› All news
Headlines
Images
rld
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
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 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
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



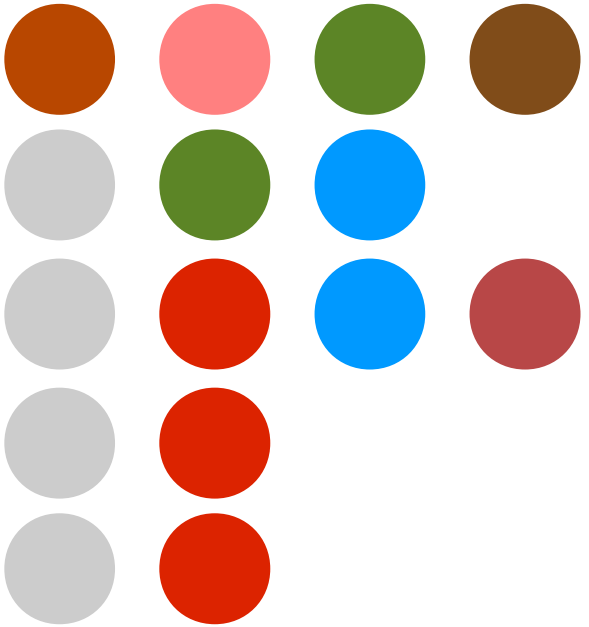


By freezelight, http://www.flickr.com/photos/63056612@N00/155554663/



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/

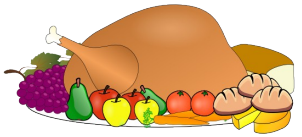
By crypto, http://www.flickr.com/photos/crypto/3201254932/sizes/l/

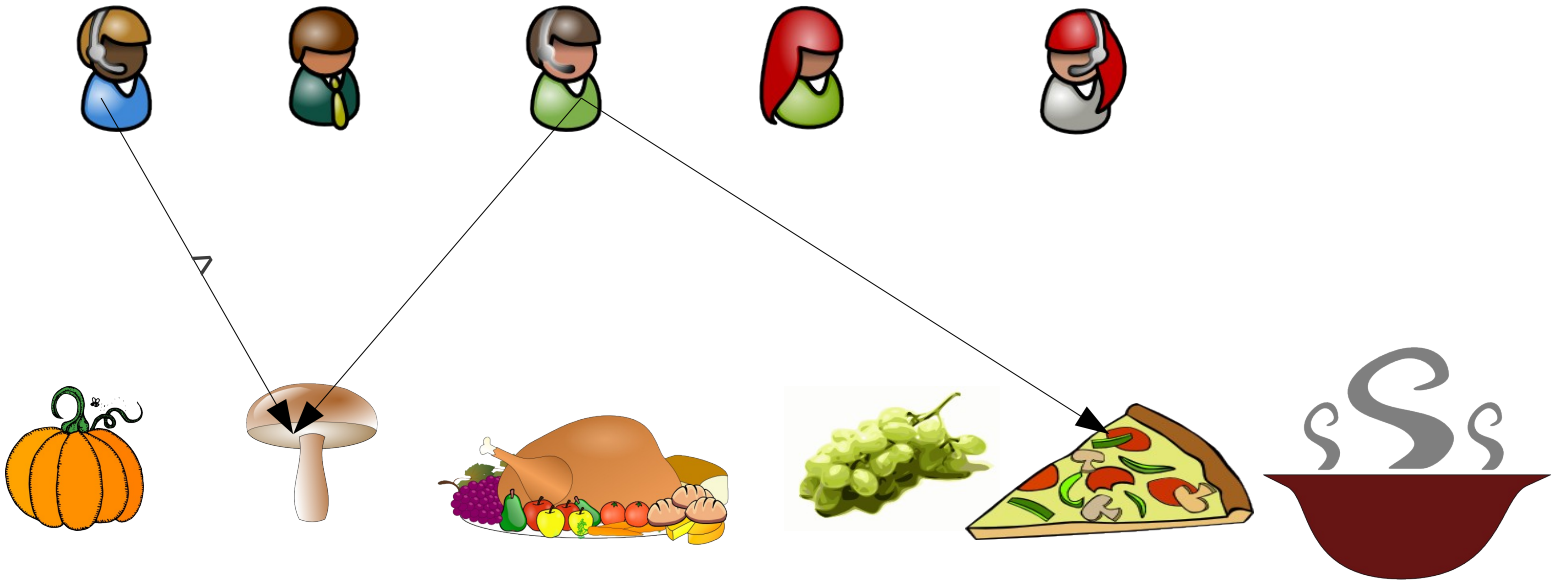
By libraryman, http://www.flickr.com/photos/libraryman/78337046/sizes/l/

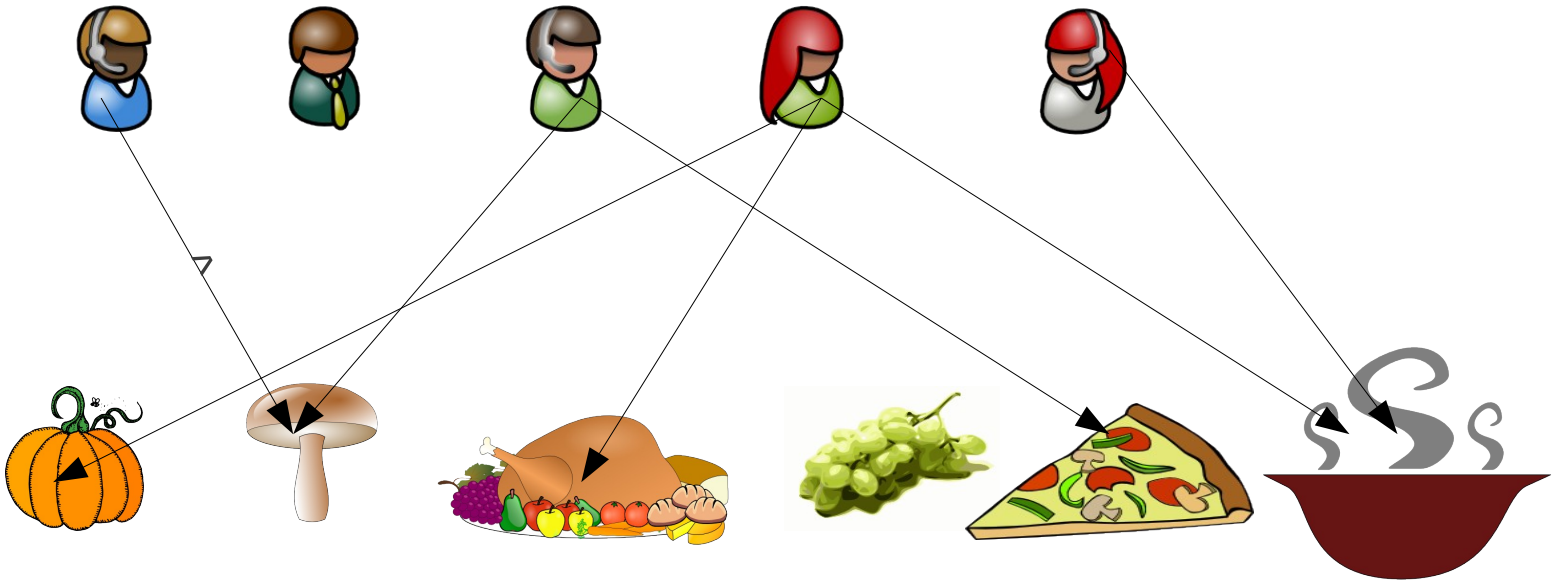


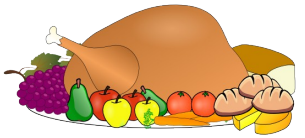


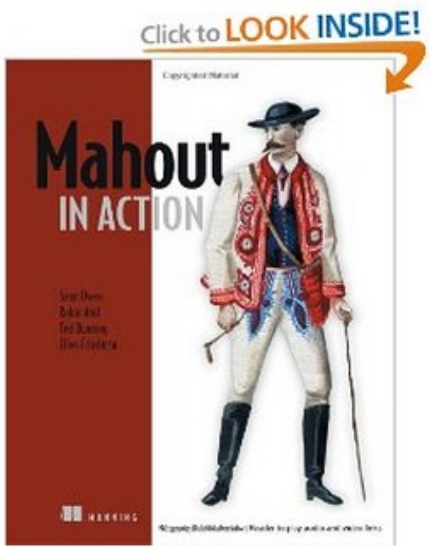






Share your own customer images

Search inside this book

Mahout in Action [Paperback]

Sean Owen (Author), Robin Anil (Author), Ted Dunning (Author), Ellen Friedman (Author)

★★★★☆ (9 customer reviews)

List Price: ~~$44.99~~

Price: $29.69 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $15.30 (34%)

In Stock.

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

Want it delivered Wednesday, November 7? Order it in the next 10 hours and 56 minutes, and choose **One-Day Shipping** at checkout. Details

Delivery may be impacted by Hurricane Sandy. Proceed to checkout to see estimated delivery dates. Learn More.

53 new from $25.64 **19 used** from $22.50

Quantity: 1

Yes, I want **FREE Two-Day Shipping** with Amazon Prime

Add to Cart

or

Sign in to turn on 1-Click ordering.

Add to Wish List

Sell Us Your Item

For a **$5.63** Gift Card

Trade in

Learn more

More Buying Choices

72 used & new from $22.50

Frequently Bought Together

Price For All Three: $82.68

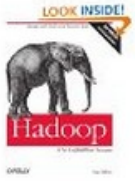
Add all three to Cart | Add all three to Wish List

Show availability and shipping details

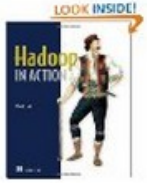
- ☑ **This item:** Mahout in Action by Sean Owen Paperback $29.69
- ☑ Hadoop in Action by Chuck Lam Paperback $25.10
- ☑ Hadoop: The Definitive Guide by Tom White Paperback $27.89

Customers Who Bought This Item Also Bought

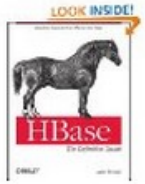
Page 1 of 17



Hadoop: The Definitive Guide
Tom White
★★★½☆ (11)
Paperback
$27.89

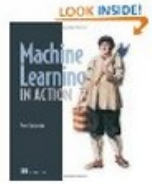


Hadoop in Action
› Chuck Lam
★★★★☆ (6)
Paperback
$25.10



HBase: The Definitive Guide
› Lars George
★★★★★ (5)
Paperback
$32.64

Lucene in Action, Second Edition: Covers Apache ...
› Michael McCandless
★★★★½ (30)
Paperback
$31.36



Machine Learning in Action
› Peter Harrington
★★★½☆ (10)
Paperback
$25.75

Programming Pig
› Alan Gates
Paperback
$33.82

http://www.flickr.com/photos/22280677@N07/2504310138

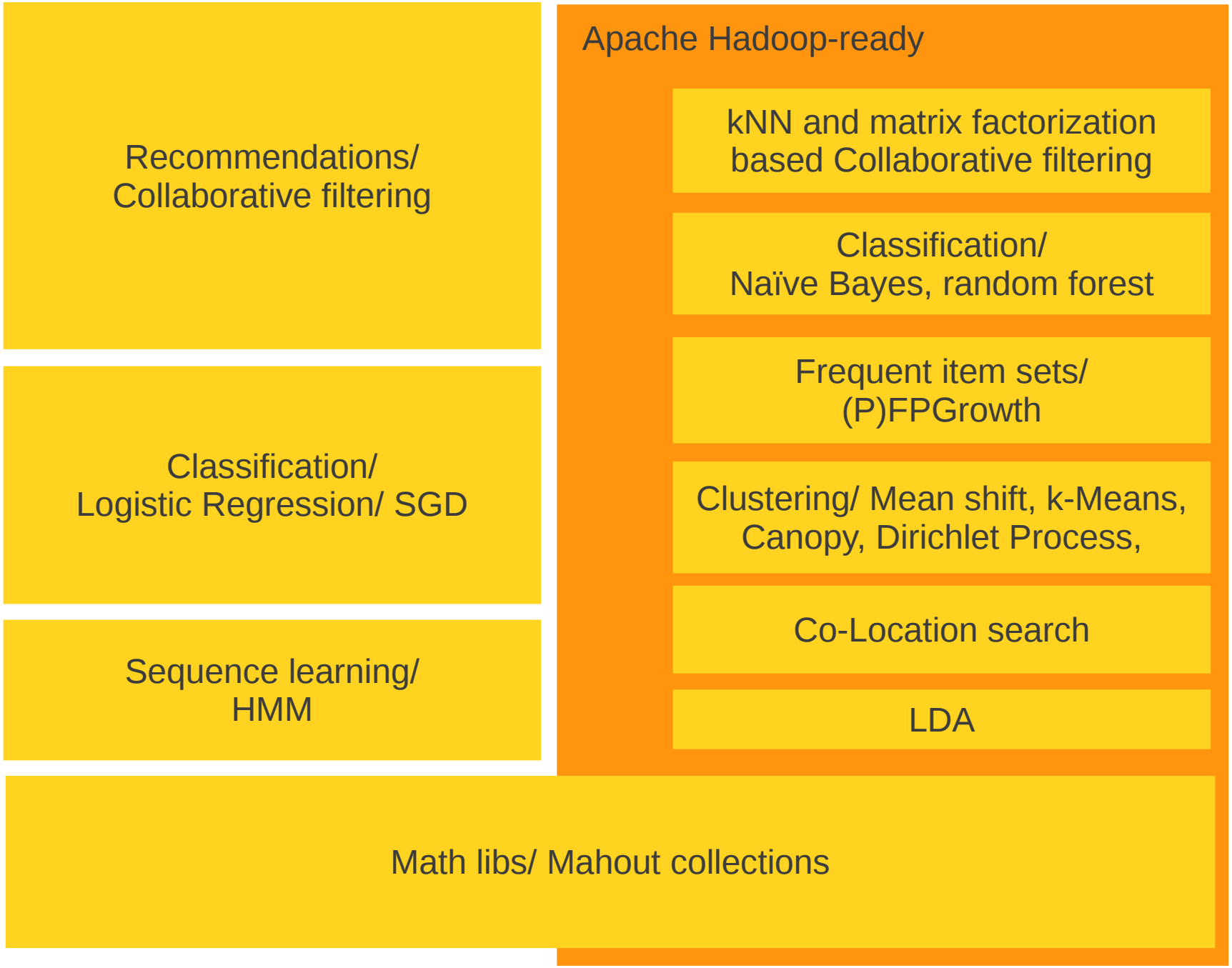
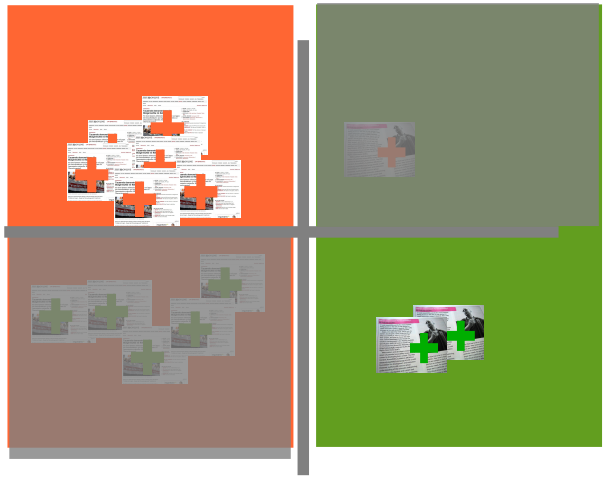

Recommendations/
Collaborative filtering
Classification/
Logistic Regression/ SGD
Sequence learning/
HMM
Apache Hadoop-ready
kNN and matrix factorization
based Collaborative filtering
Classification/
Naïve Bayes, random forest
Frequent item sets/
(P)FPGrowth
Clustering/ Mean shift, k-Means,
Canopy, Dirichlet Process,
Co-Location search
LDA
Math libs/ Mahout collections


Image by ZapTheDingbat (Light meter)
http://www.flickr.com/photos/zapthedingbat/3028168415
DIN
LENINGRAD

Accuracy

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

- Problems:
  - What if class distribution is skewed?



Precision/ Recall

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

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

- Problem:
  - Depends on decision threshold.

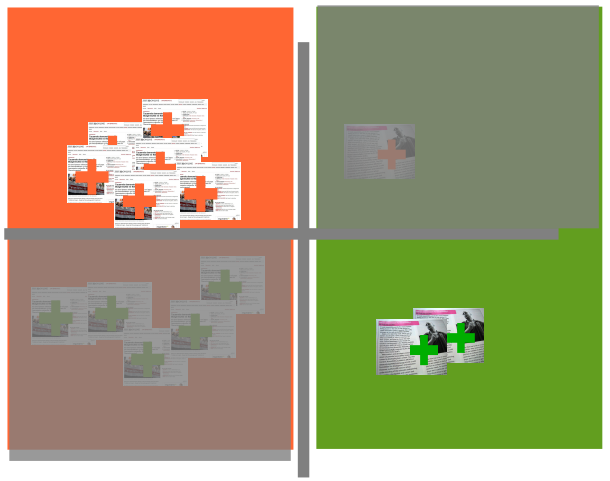
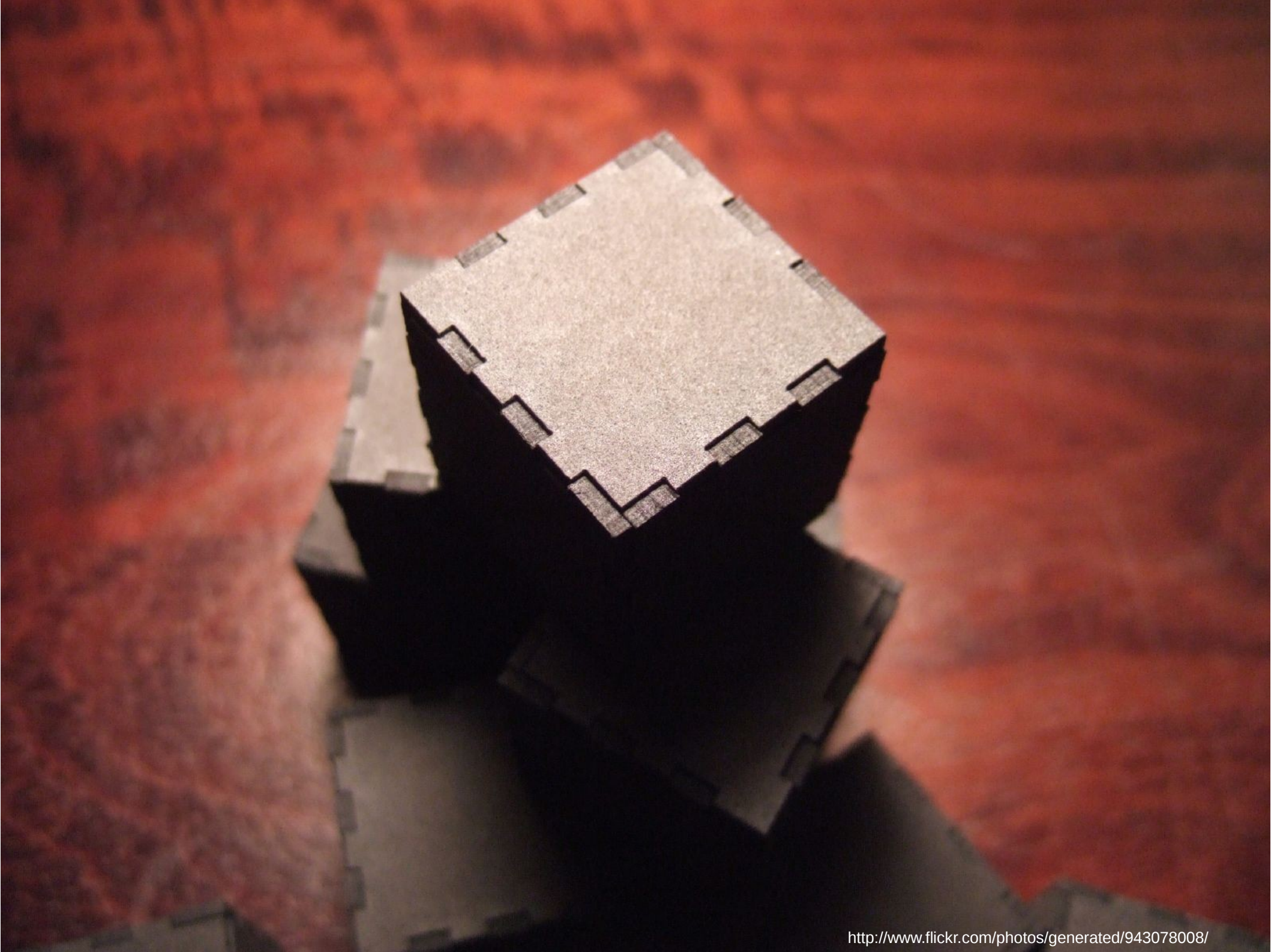


Image taken by wwarby.
http://www.flickr.com/photos/wwarby/3297205226/


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


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

Libraries to have a look at:

Vowpal Wabbit	Mallet
LibSvm	LibLinear
Libfm	Incanter
GraphLab	Skikits learn

Frameworks worth mentioning:

Apache Mahout	Apache Giraph
Matlab / Otave	R
Shogun	Weka
RapidI	MyMedialight

Where to get more information:

"Mahout in Action" - Manning
"Taming Text" - Manning
"Machine Learning" - Andrew Ng

https://cwiki.apache.org/confluence/display/MAHOUT/Books+Tutorials+and+Talks

https://cwiki.apache.org/confluence/display/MAHOUT/Reference+Reading

Get your hands dirty:

http://kaggle.com

https://cwiki.apache.org/confluence/display/MAHOUT/Collections

Where to meet these people:

RecSys	ICML
NIPS	ECML
KDD	WSDM
PKDD	JMLR

Mahout is general purpose – add your domain knowledge.

Understand your data.

Combine several tools to match your problem.


Get started today with the right tools.
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

WARNING

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



Paritosh Ranjan
Dmitriy Lyubimov
Shannon Quinn
Sebastian Schelter
Jake Mannix
Benson Margulies
Robin Anil
Sean Owen
Grant Ingersoll
Drew Farris
Jeff Eastman
Ted Dunning
Isabel Drost

Become a committer: Of Apache Mahout


Emeritus:

Niranjan Balasubramanian
Otis Gospodnetic
David Hall
Erik Hatcher
Ozgur Yilmazel
Dawid Weiss
Karl Wettin
AbdelHakim Deneche

http://BerlinBuzzwords.de – June 2013 in Berlin/ Germany.

Online – user@mahout.apache.org and dev@mahout.apache.org

Image by: Patrick McEvoy

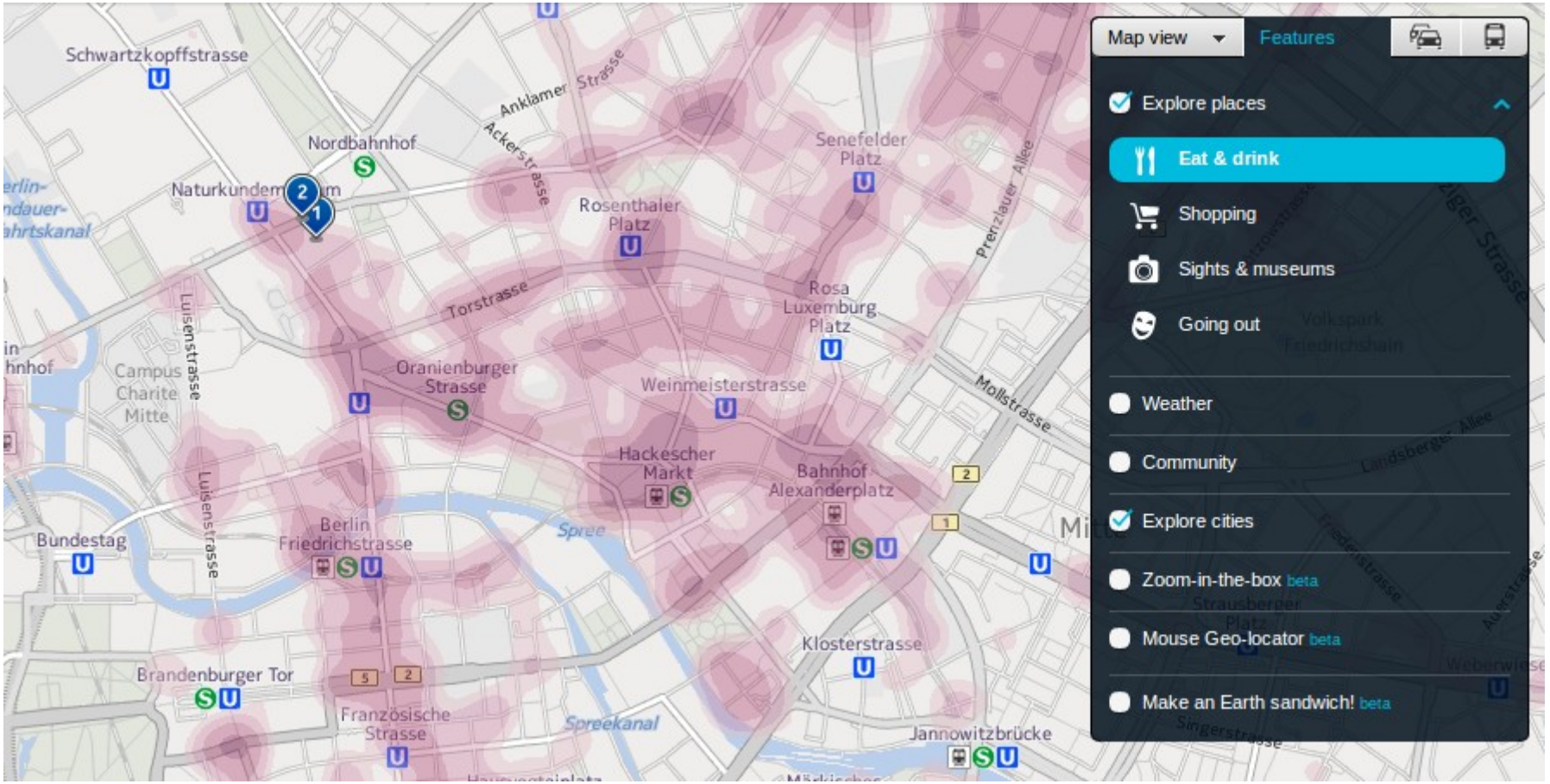
Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.


Map view
Features
Explore places
Eat & drink
Shopping
Sights & museums
Going out
Weather
Community
Explore cities
Zoom-in-the-box beta
Mouse Geo-locator beta
Make an Earth sandwich! beta
Schwartzkopffstrasse
Nordbahnhof
Anklamer Strasse
Ackerstrasse
Senefelder Platz
Rosenthaler Platz
Prenzlauer Allee
Torstrasse
Rosa Luxemburg Platz
Oranienburger Strasse
Weinmeisterstrasse
Mollstrasse
Hackescher Markt
Bahnhof Alexanderplatz
Campus Charite Mitte
Luisenstrasse
Bundestag
Berlin Friedrichstrasse
Spree
Klosterstrasse
Brandenburger Tor
Französische Strasse
Spreekanal
Jannowitzbrücke





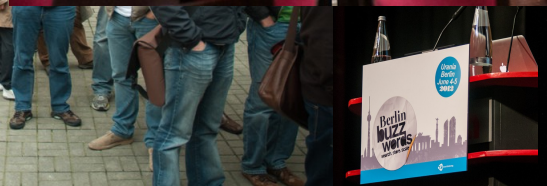
Kasse 1










Berlin buzzwords


http://BerlinBuzzwords.de – June 2013 in Berlin/ Germany.

Online – user@mahout.apache.org and 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.