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

Hello Apache Con!

Hello Apache Con!

Hello Apache Con!

Hello Apache Con!

Machine learning
background?

Hello Apache Con!

Hello Apache Con!

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?

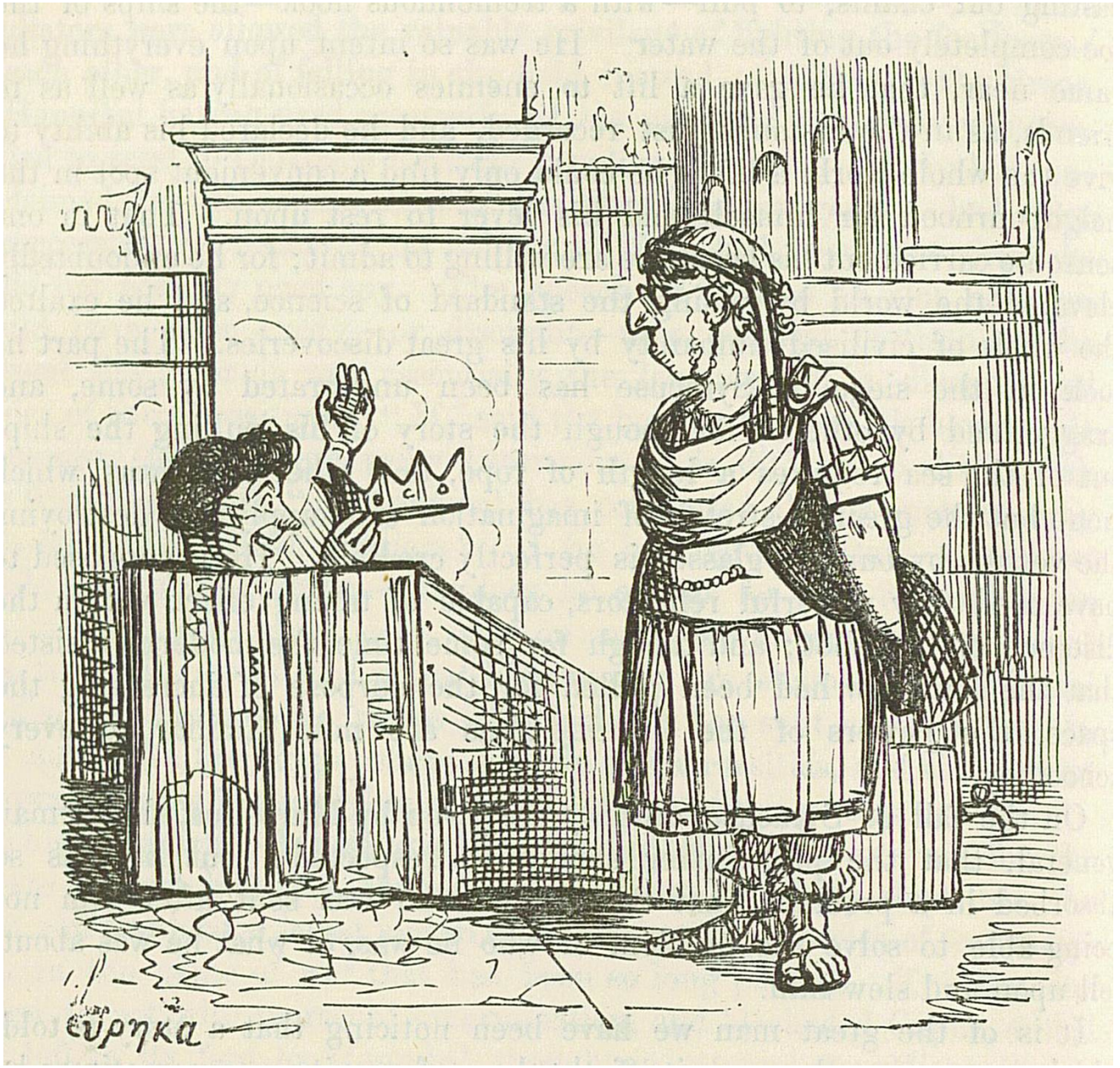


Image by John Leech, from: The Comic History of Rome by Gilbert Abbott A Beckett.
Bradbury, Evans & Co, London, 1850s
Archimedes taking a Warm Bath

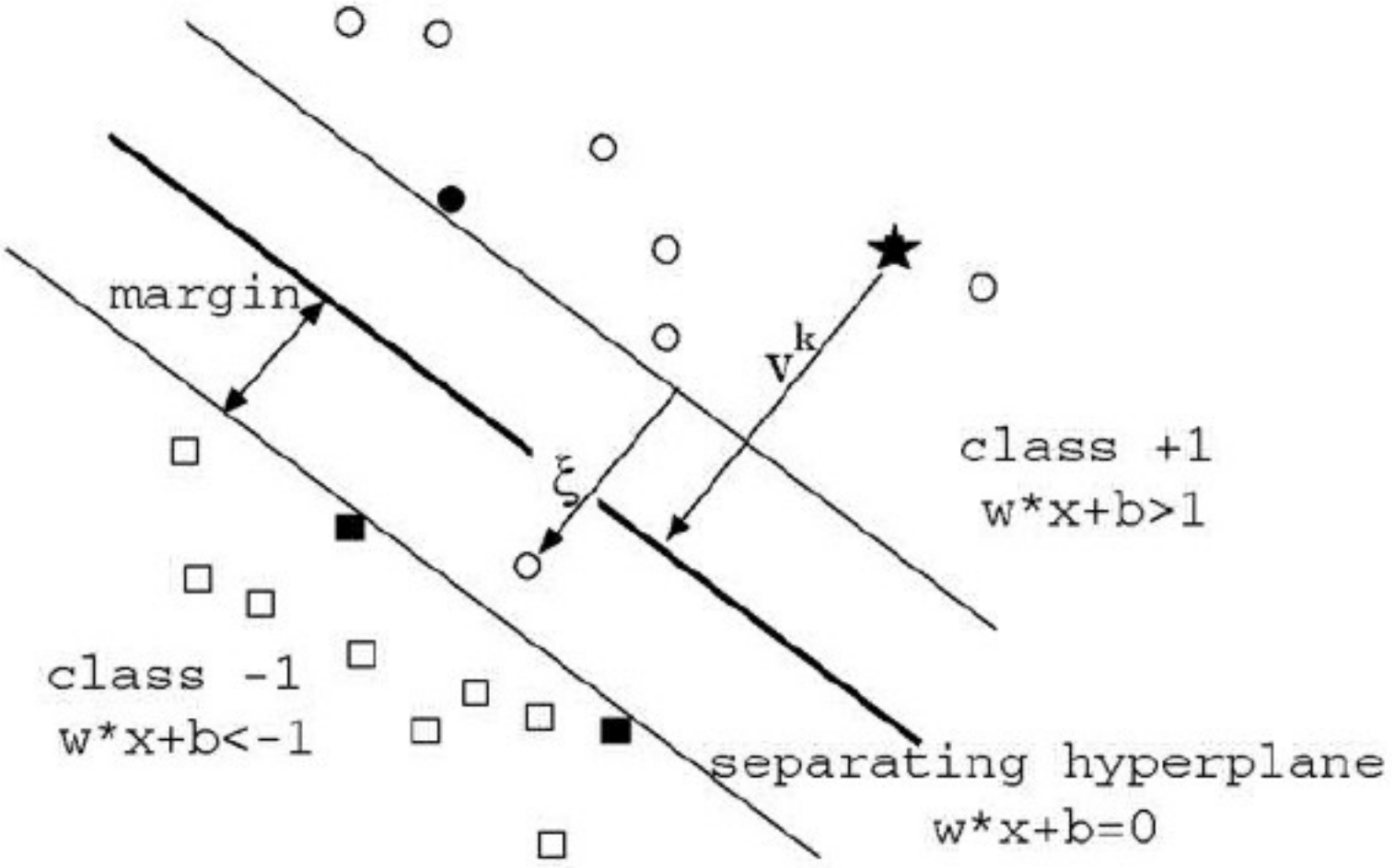
Archimedes model of nature

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

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


June 25, 2008 by chase-me
http://www.flickr.com/photos/sasy/2609508999
NICHOLAS SPARKS
THE NOTEBOOK
Nicholas Sparks
THE DEVIL WEARS PRADA
Best Friends
Maya Angelou
The Heart of a Woman
comeback
CHRISTOPHER ANDERSEN
AMERICAN EVITA

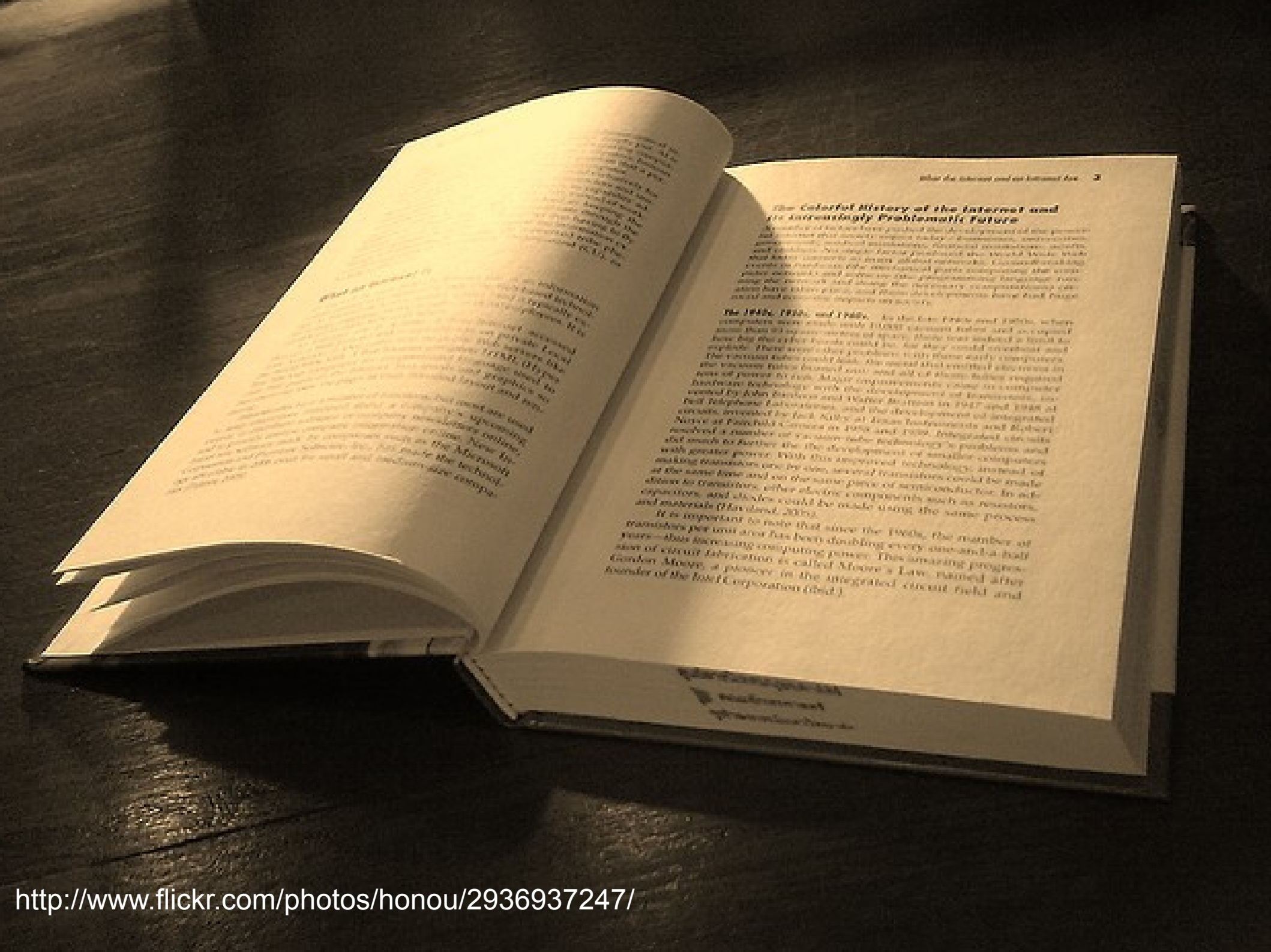
An SVM's model of nature



The challenge

Mission

Provide scalable data mining algorithms.



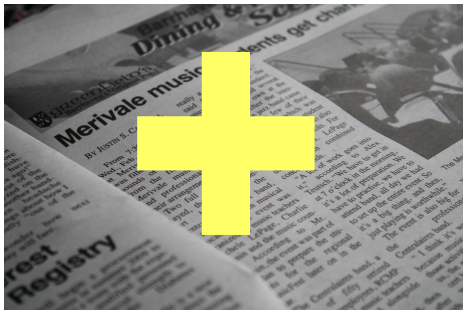
http://www.flickr.com/photos/honou/2936937247/

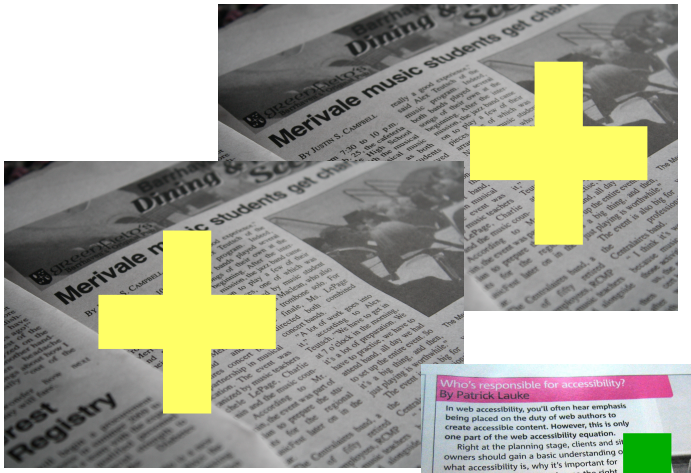
HowTo: From data to information.

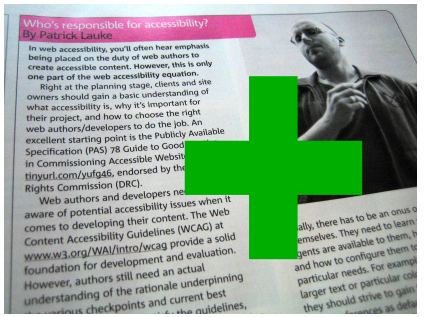
January 3, 2006 by Matt Callow
http://www.flickr.com/photos/blackcustard/81680010

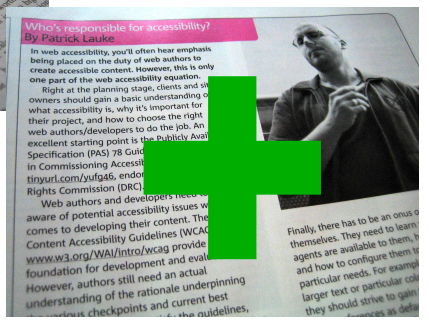
http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

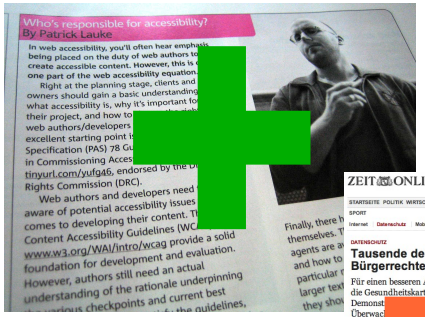
http://www.flickr.com/photos/redux/409356158/

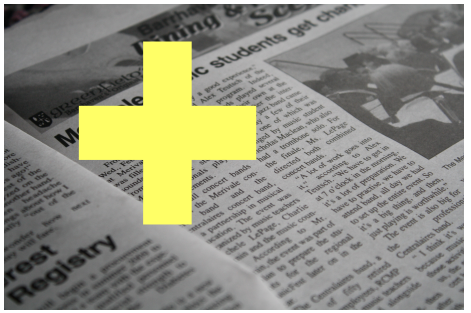


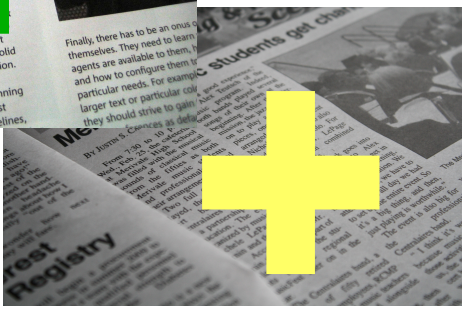






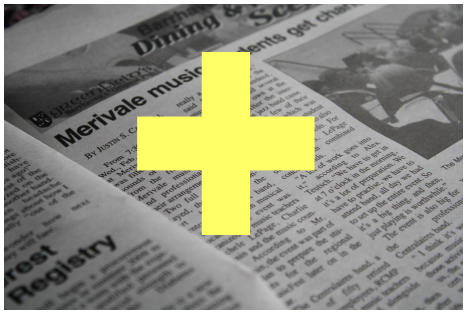


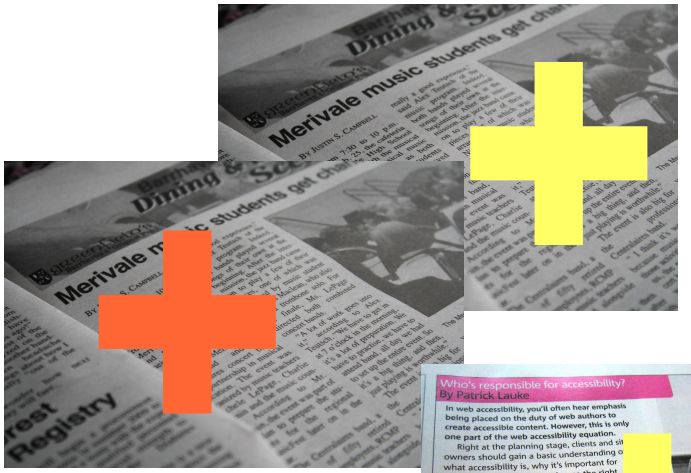


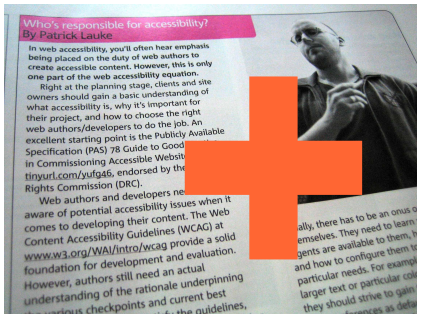


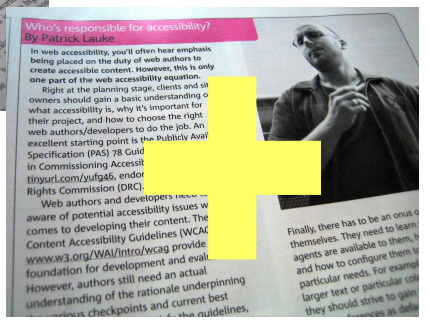
http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

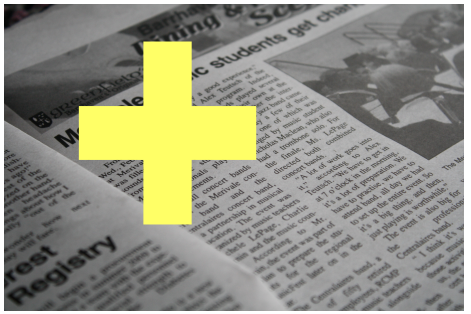
http://www.flickr.com/photos/redux/409356158/

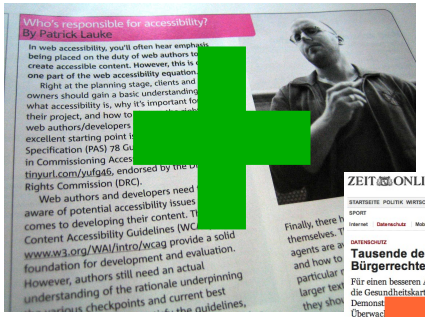


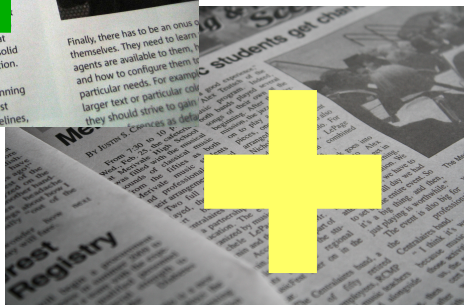










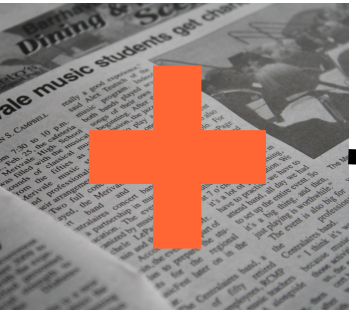


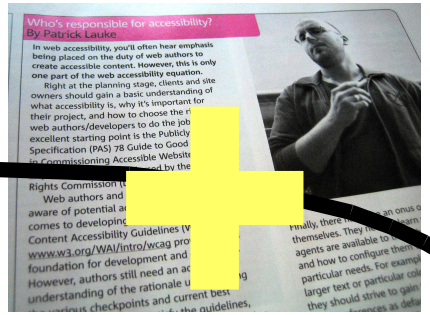
http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

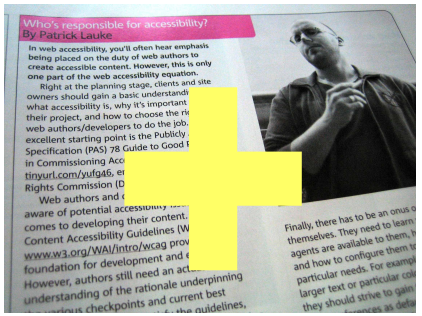
http://www.flickr.com/photos/redux/409356158/

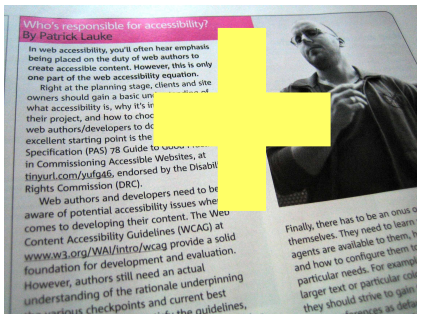
http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

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









http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

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

http://www.flickr.com/photos/topsy/204929063/

ZEIT ONLINE | DATENSCHUTZ

Partnersuche

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIU
SPORT

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.

© Tim Brakemeier/dpa

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.

DATUM
QUELL
KOMM
EMPFE
Buzz
ARTIK
SCHLA
Datens

NEU IM R
1. DATENS im Netz
2. FUTURE
3. URHEB gut"
4. SPAM A den Hä
5. IPHONE

NEU AUF
1. NACHR
2. CDU Un
3. GEDEN Ressen
4. BUNDESLIGA Der BVB und die Freiburger siegen
5. DAAD Das Ende der Ära Bode

ANZEIGE

experteer.de

Merivale mus

By Justin S. Campbell

From 7:30 to 10 p.m. Wed., Feb. 25, the cafeter Merivale High Sc filled with the of classic fifti

Who's respons
By Patrick Lauke

new China

although at times ter fying in their aggress scale, they also ref the country's effor give shape to an en ing national identi Foster's airport nal, the world's is the purest exp of China's emb the Modernist swooping form suggests two angs placed si has been com dragon. Yet cedent is Airport monument conceived Speer in gateway

Tim Griffith/PTW Architects

ent ceremonial axis.

The National Aquatics Center lies on the

among the most imaginative architec- tural feats in recent memory. Critics have incessantly described profile projects as bullish the nation's budding buildings are

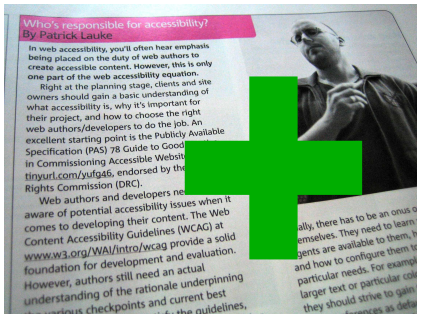
Europe. Both are part bile society that exte Grand Central Term the great train halls Like Tempelhof, nal boasts a swe evokes the glam closing a surp

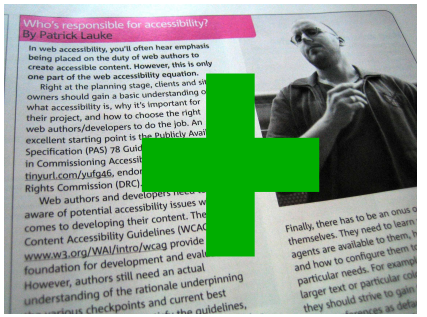
http://www.flickr.com/photos/disowned/1158260369/

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**.

http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

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





From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.

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.

© Tim Brakemeier/dpa

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.

Die Demonstranten wandten sich unter anderem gegen die Volkszählung

DATUM 11.9.2010 - 17:20 Uhr
QUELLE ZEIT ONLINE, AFP, dpa
KOMMENTARE 4
EMPFEHLEN E-Mail verschicken | Facebook, Twitter, Buzz ...
ARTIKEL DRUCKEN Druckversion | PDF
SCHLAGWORTE Datenschutz | Demonstration | Datensicherheit | Medienpolitik

NEU IM RESSORT

1. 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"
4. 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

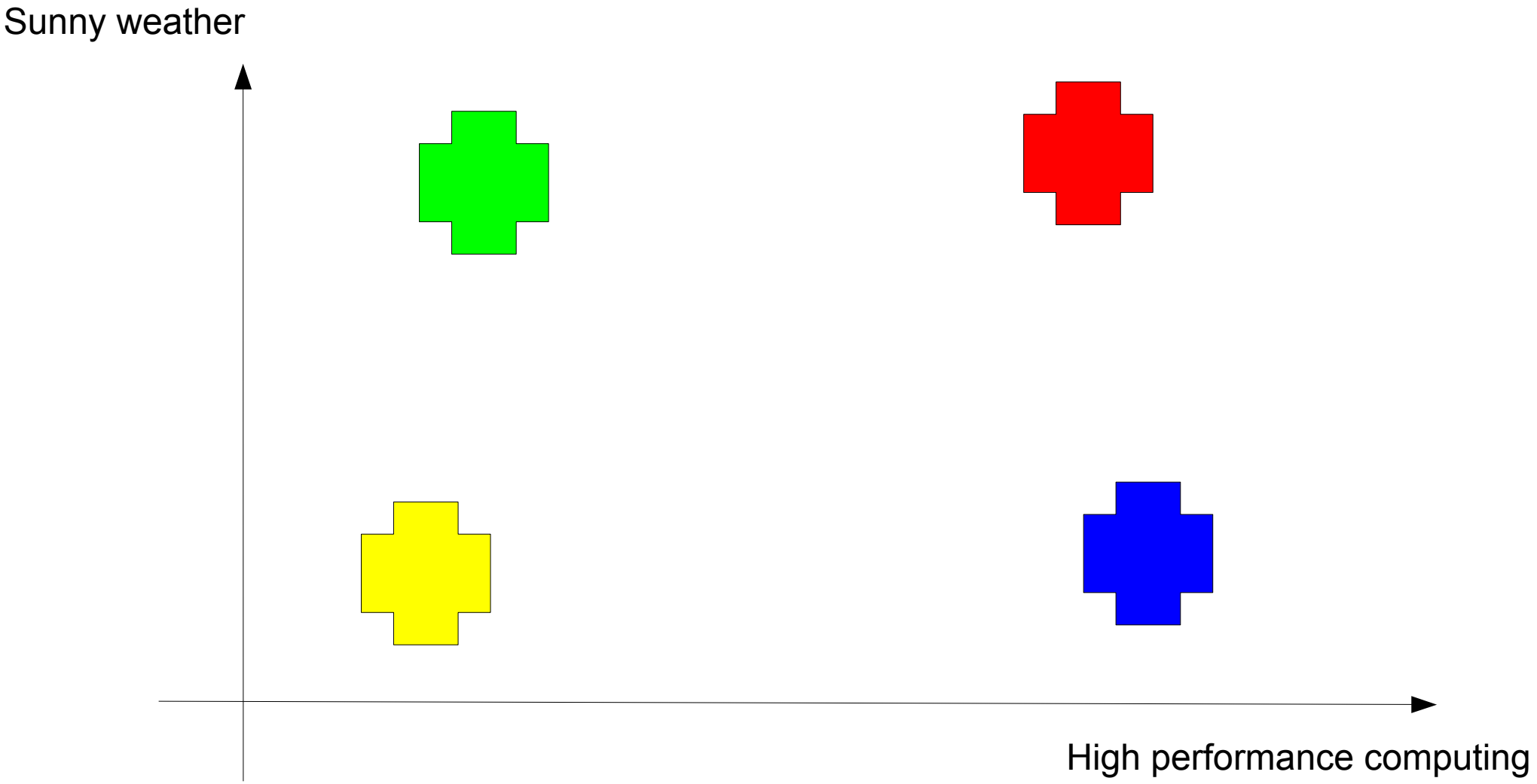
1. NACHRUF Die Freie - Bärbel Bohley ist tot
2. CDU Union streitet über konservatives Profil
3. GEDENKTAG 9/11 Obama warnt vor religiösen Ressentiments
4. BUNDESLIGA Der BVB und die Freiburger siegen
5. DAAD Das Ende der Ära Bode

- Remove noise.

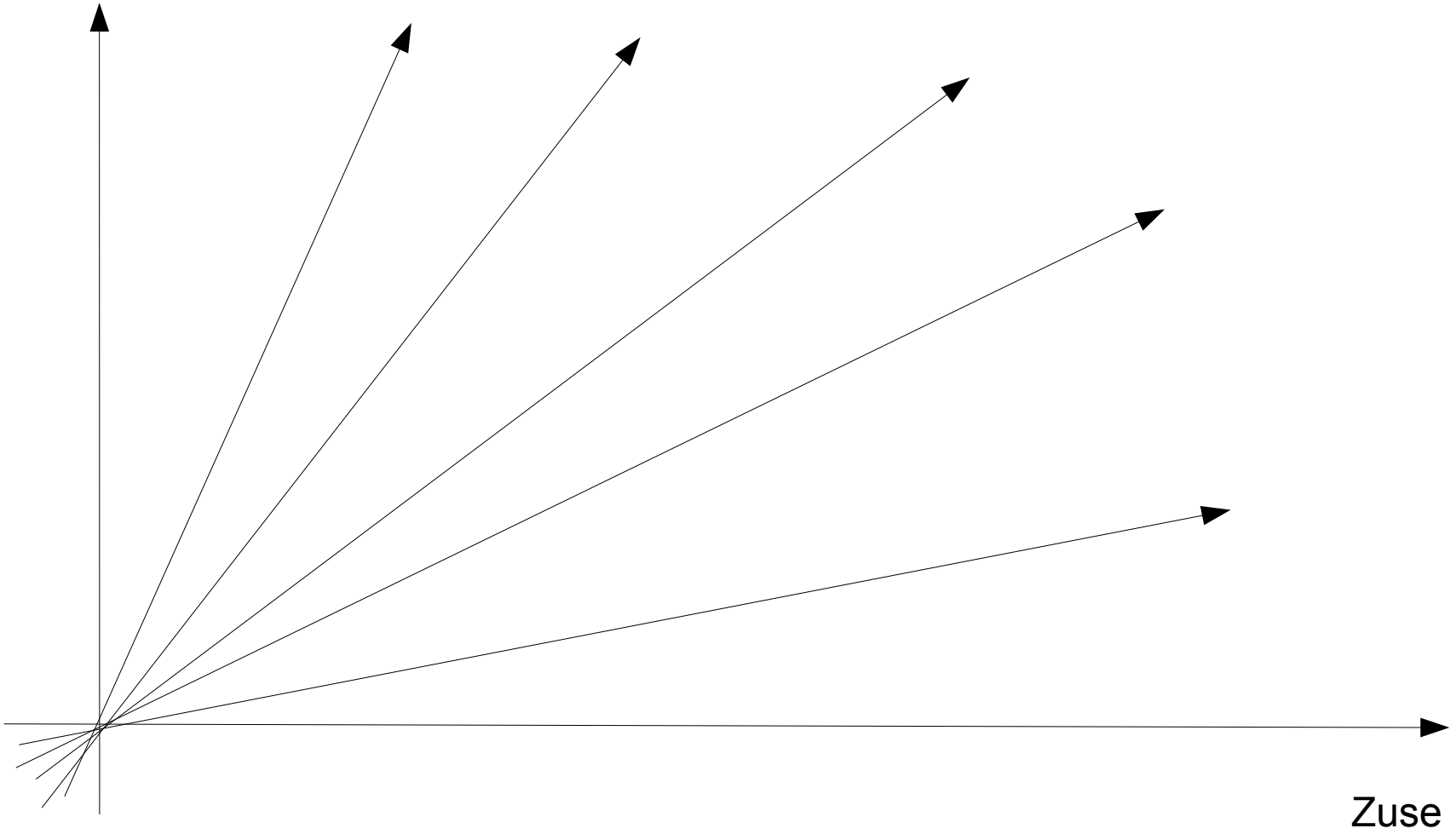
- Remove noise.
- Convert text to vectors.

From texts to vectors

If we looked at two words only:



Aaron



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.
- Problem:
  - Long texts get larger values.

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

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.
- Problem:
  - Additional domain knowledge ignored.

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

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.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.

Algorithm choice

- Naive Bayes.
  - Cannot reliably indicate how certain its classification is.

Algorithm choice

- Naive Bayes.
  - Cannot reliably indicate how certain its classification is.
- Logistic Regression.
- Complement. NB.
- Random Forests.

Algorithm choice

- Do you
  - want to interpret the resulting model?
  - want to update the model in an online fashion?

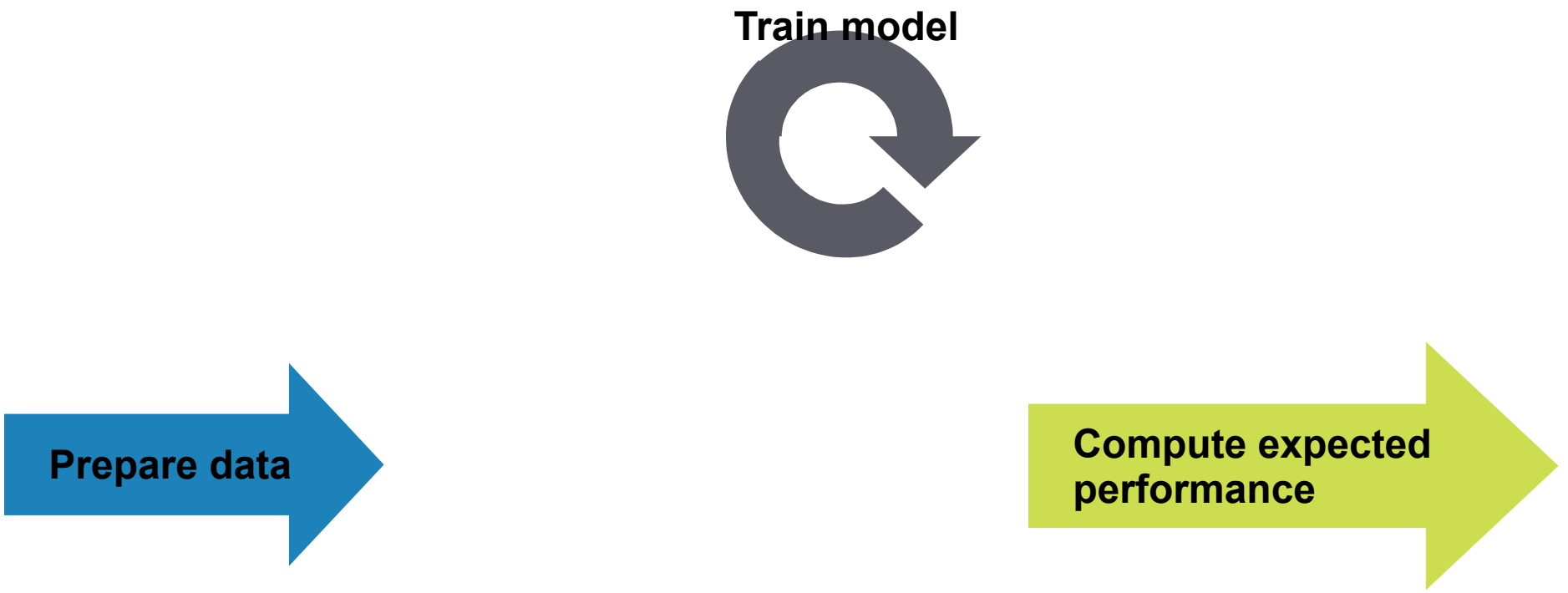
- The data you are working with
  - lives in high-dim feature space but is sparse?
  - has features that might depend on each other?
  - has outliers?
  - has missing values?

From data to information.

- Collect data and define your learning problem.
- Data preparation.
- Training a prediction model.
- Checking the performance of your model.

Goals

- Did I use the best model parameters?
- How well will my model perform in the wild?


Train model
Prepare data
Compute expected performance

Performance

- Use same data for training and testing.

- Problem:
  - Highly optimistic.
  - Model generalization unknown.

Performance

- Use same data for training and testing.

DON'T

- Problem:
  - Highly optimistic.
  - Model generalization unknown.

Performance

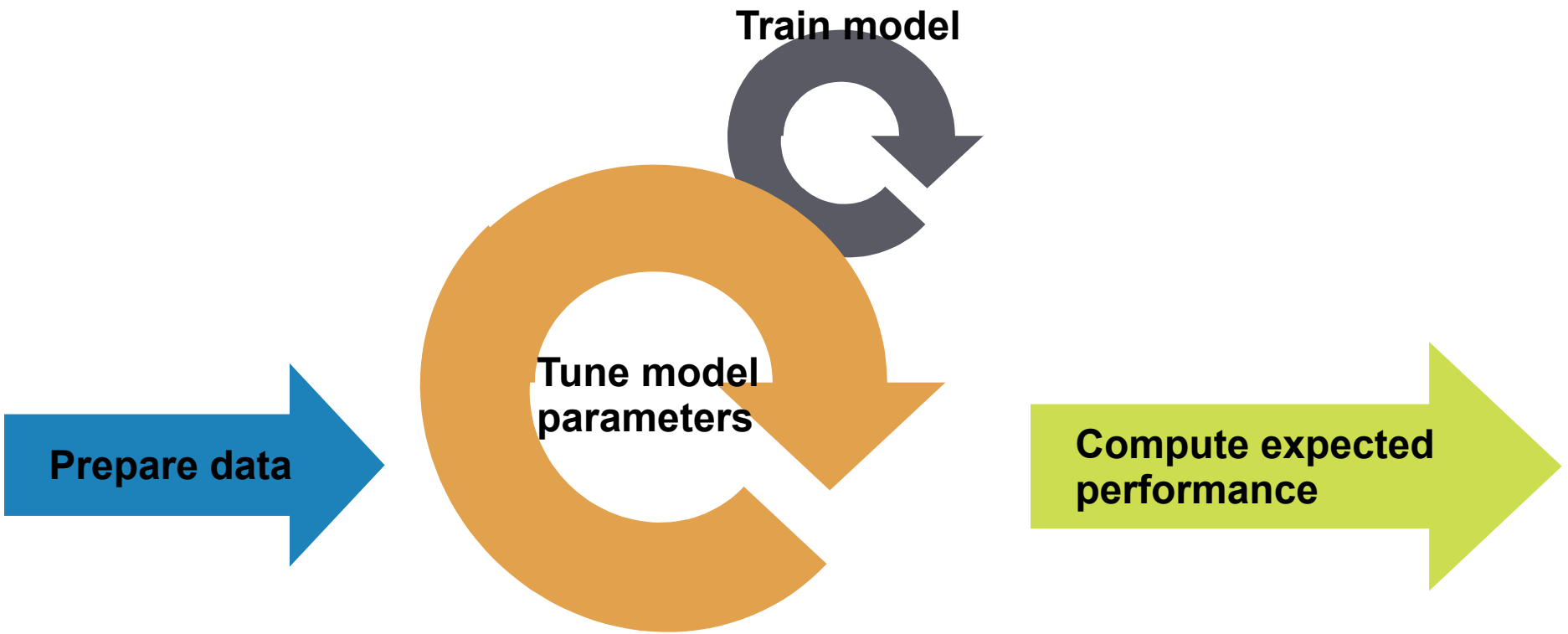
- Use just a fraction for training.
- Set some data aside for testing.

- Problems:
  - Pessimistic predictor: Not all data used for training.
  - Result may depend on which data was set aside.

Performance

- Partition your data into n fractions.
- Each fraction set aside for testing in turn.

- Problem:
  - Still a pessimistic predictor.

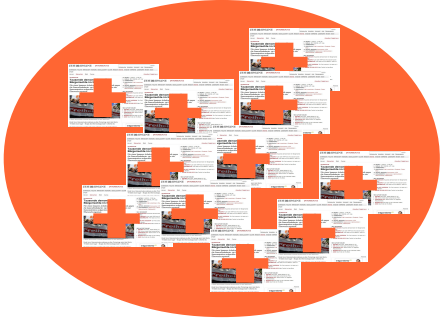

Train model
Tune model parameters
Prepare data
Compute expected performance

Performance

- Use just a fraction for training.
- Set some data aside for tuning and testing.

- Problems:
  - Highly optimistic.
  - Parameters manually tuned to testing data.

Performance



- Use just a fraction for training.
- Set some data aside for validation and testing.

DON'T

- Problems:
  - Highly optimistic.
  - Parameters manually tuned to testing data.

Performance

- Use just a fraction for training.
- Set some data aside for tuning.
- Set another set of data aside for testing.

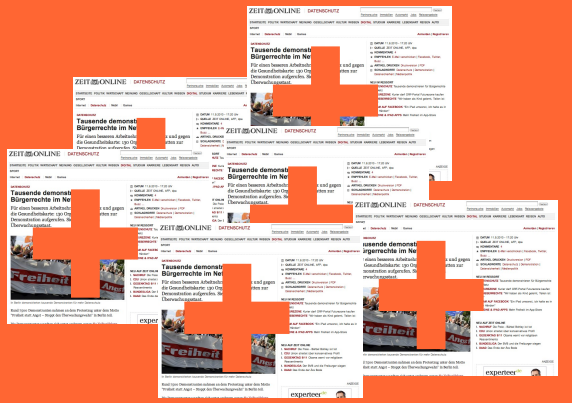
- Problems:
  - Pretty pessimistic as not all data is used.
  - May depend on which data was set aside.

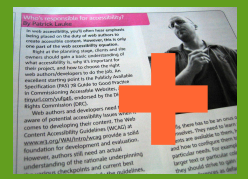
Performance Measures

Correct prediction: Orange

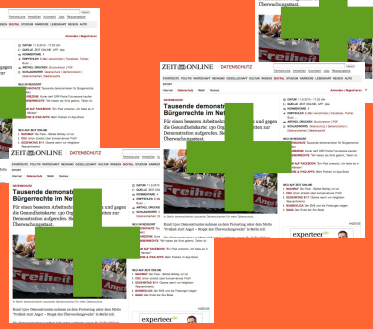
Correct prediction: Green

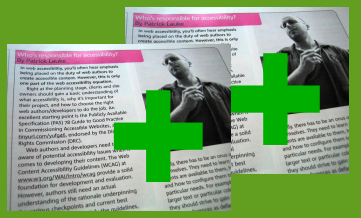
Model prediction: Orange





Model prediction: Green

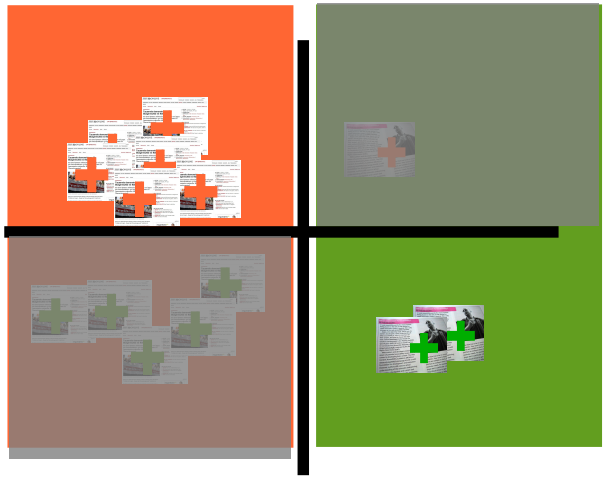




Accuracy

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

- Problems:
  - What if class distribution is skewed?

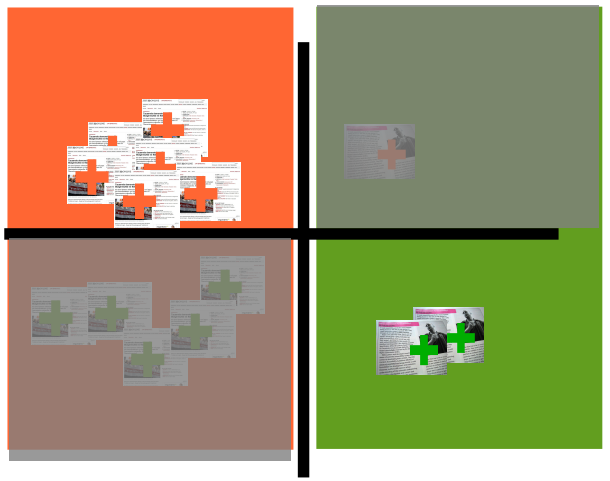


Precision/ Recall

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

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

- Problem:
  - Depends on decision threshold.

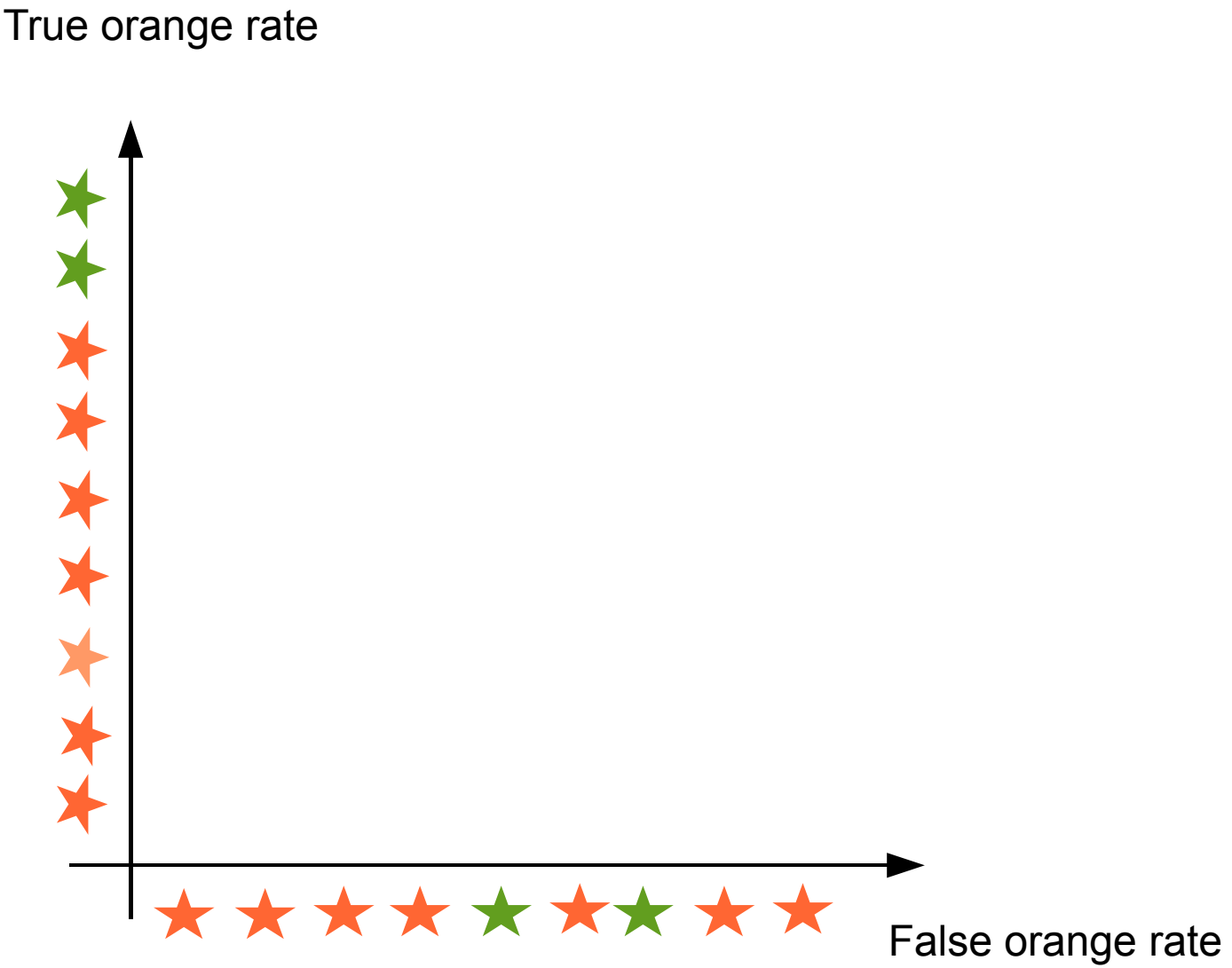


ROC Curves

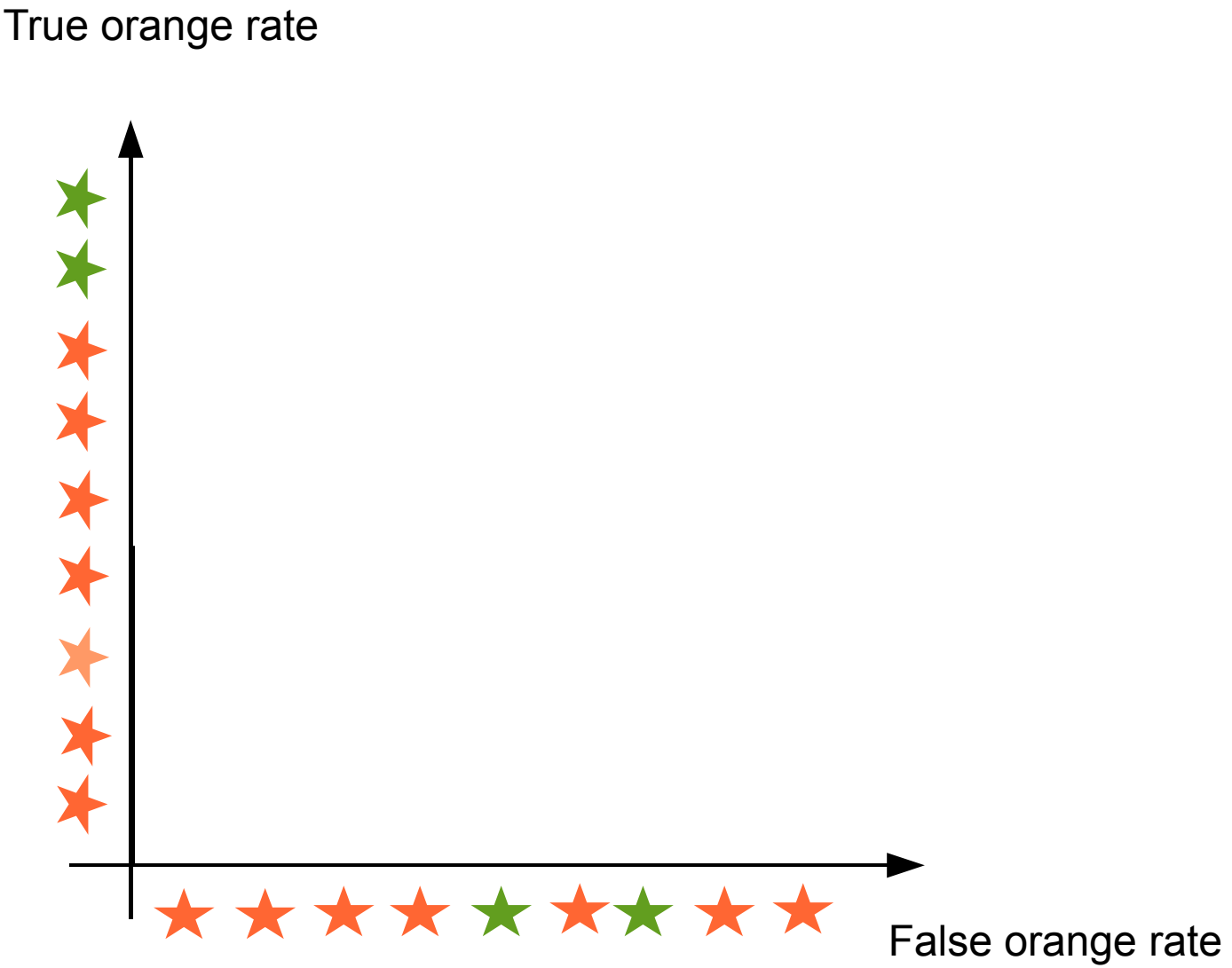
★ ★ ★ ★ ★ ★ ★ ★ ★

ROC Curves

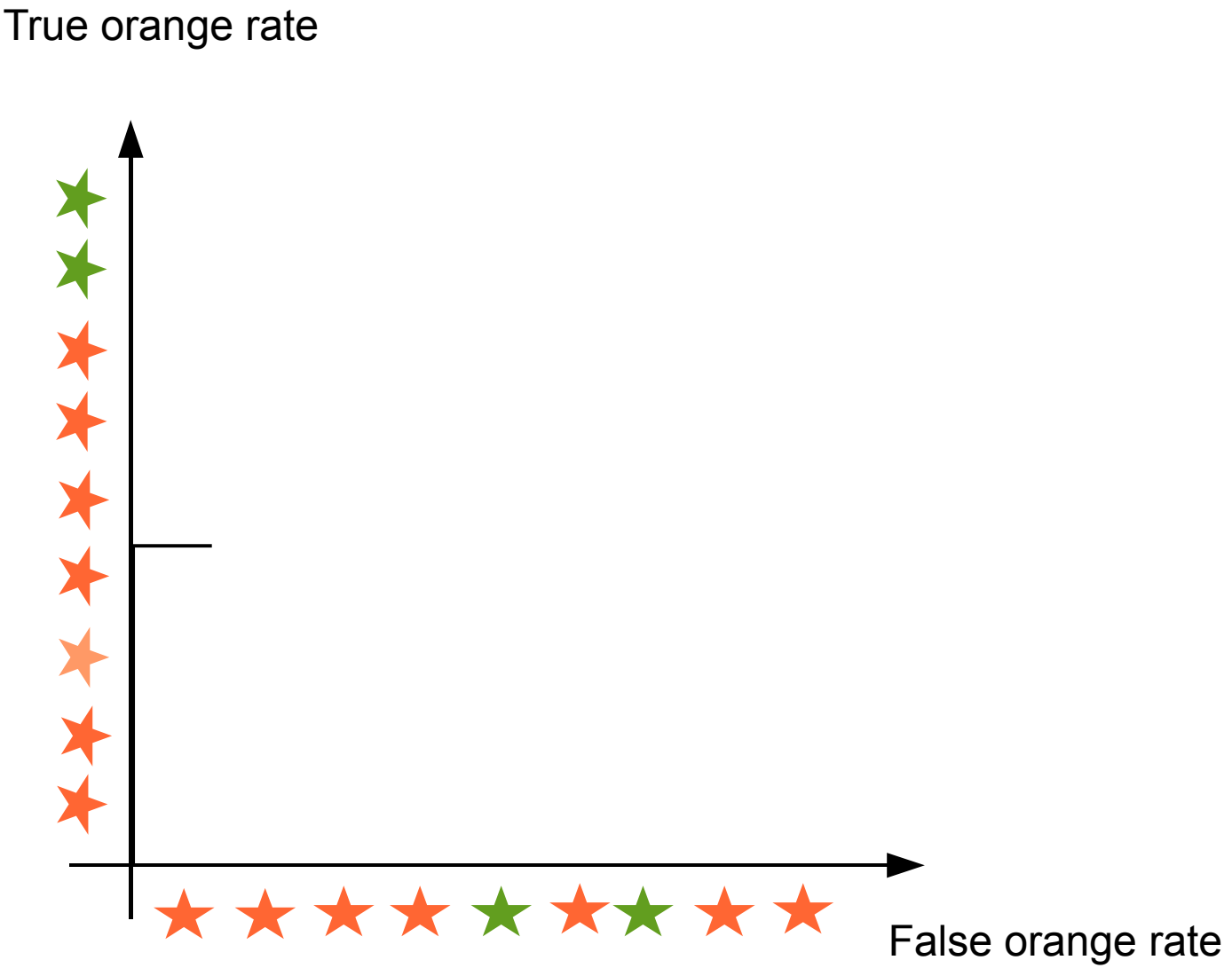
ROC Curves



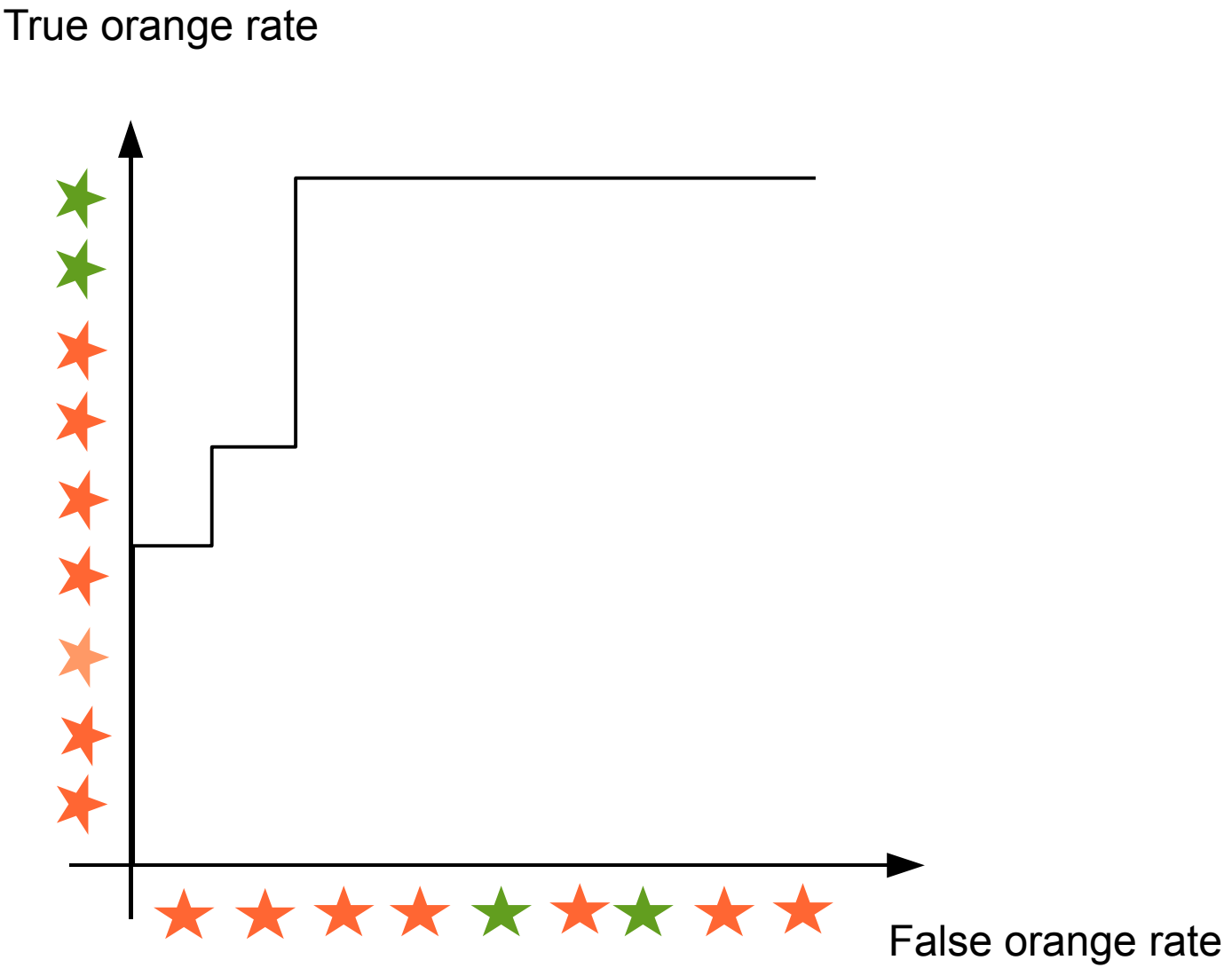
ROC Curves



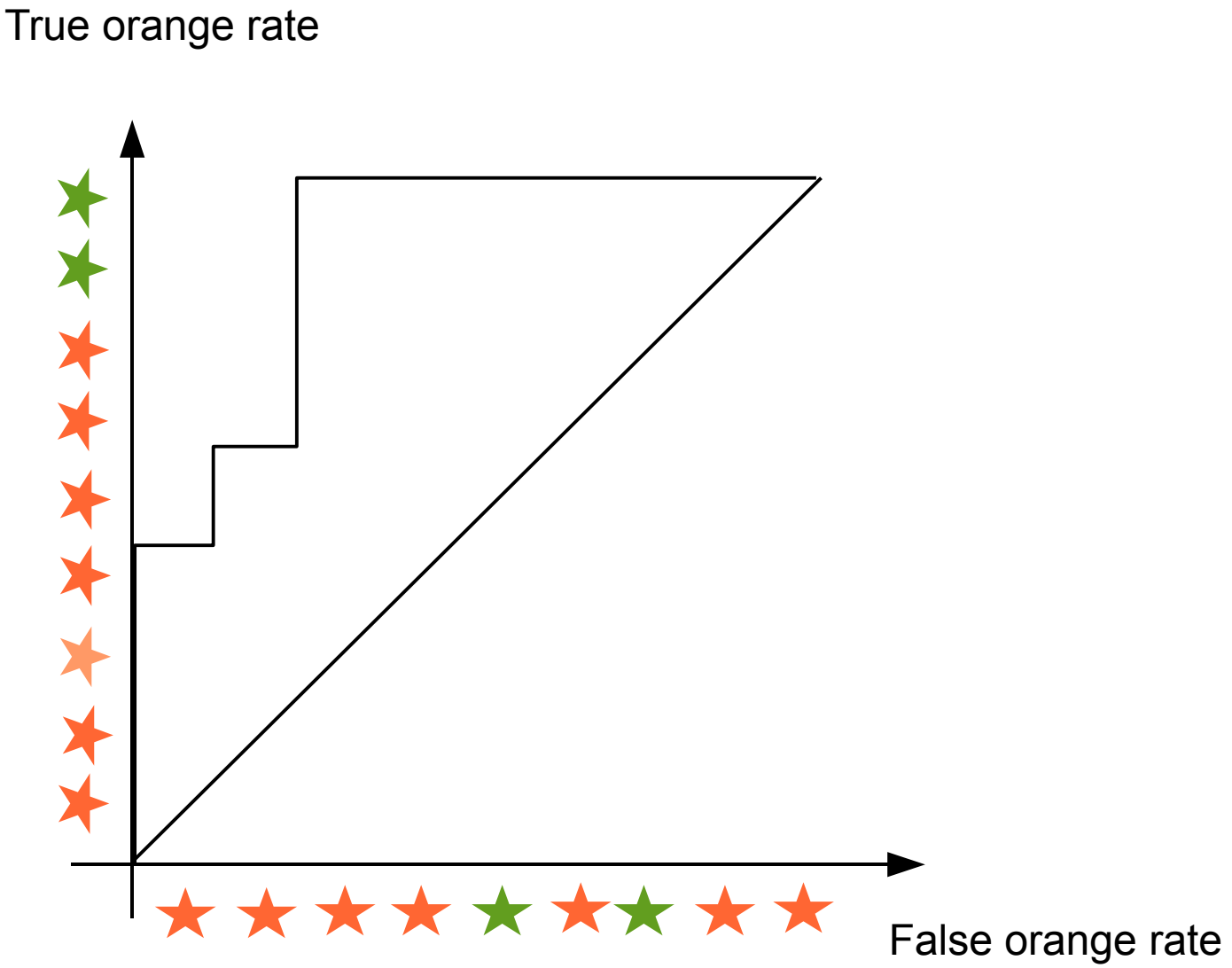
ROC Curves



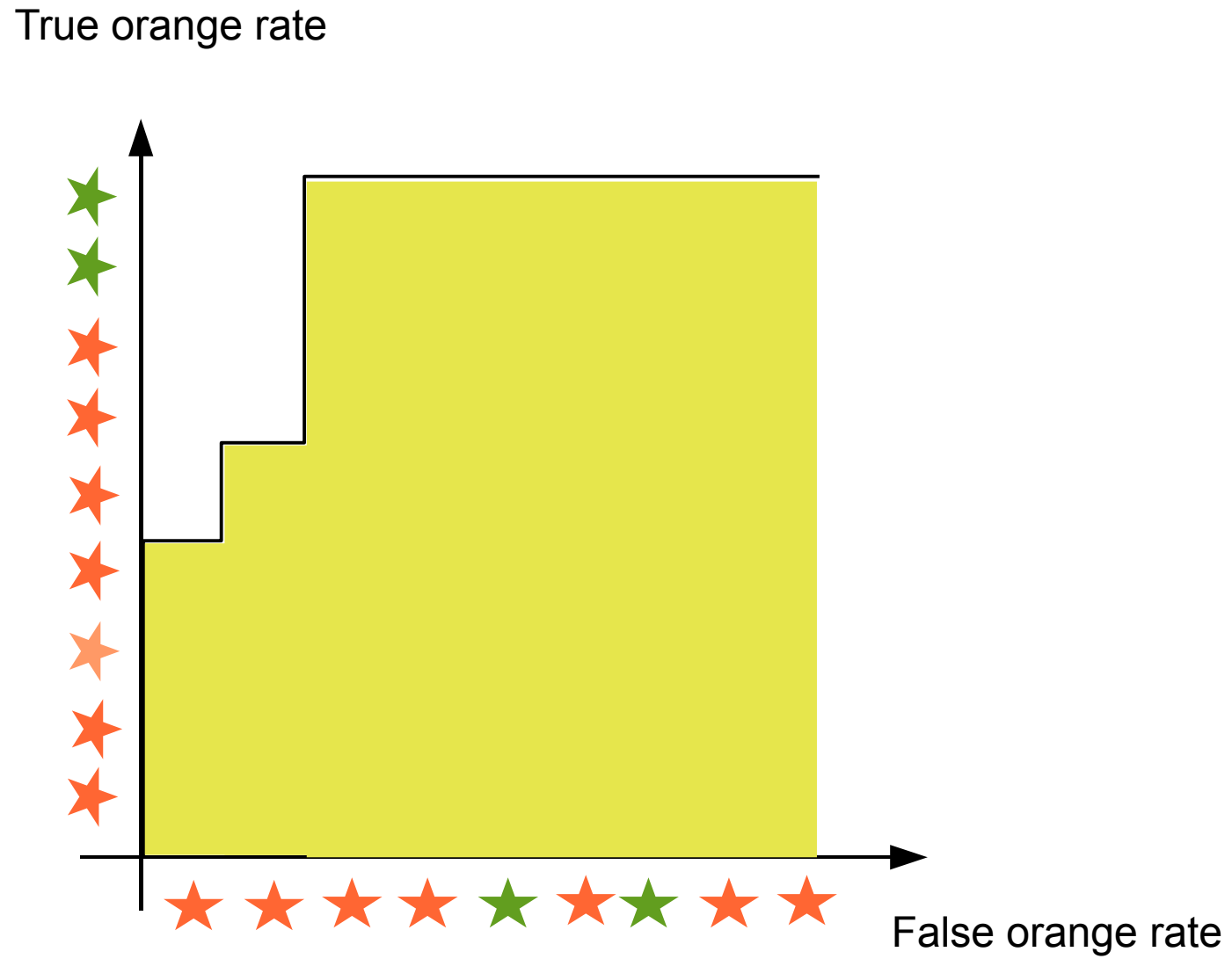
ROC Curves



ROC Curves

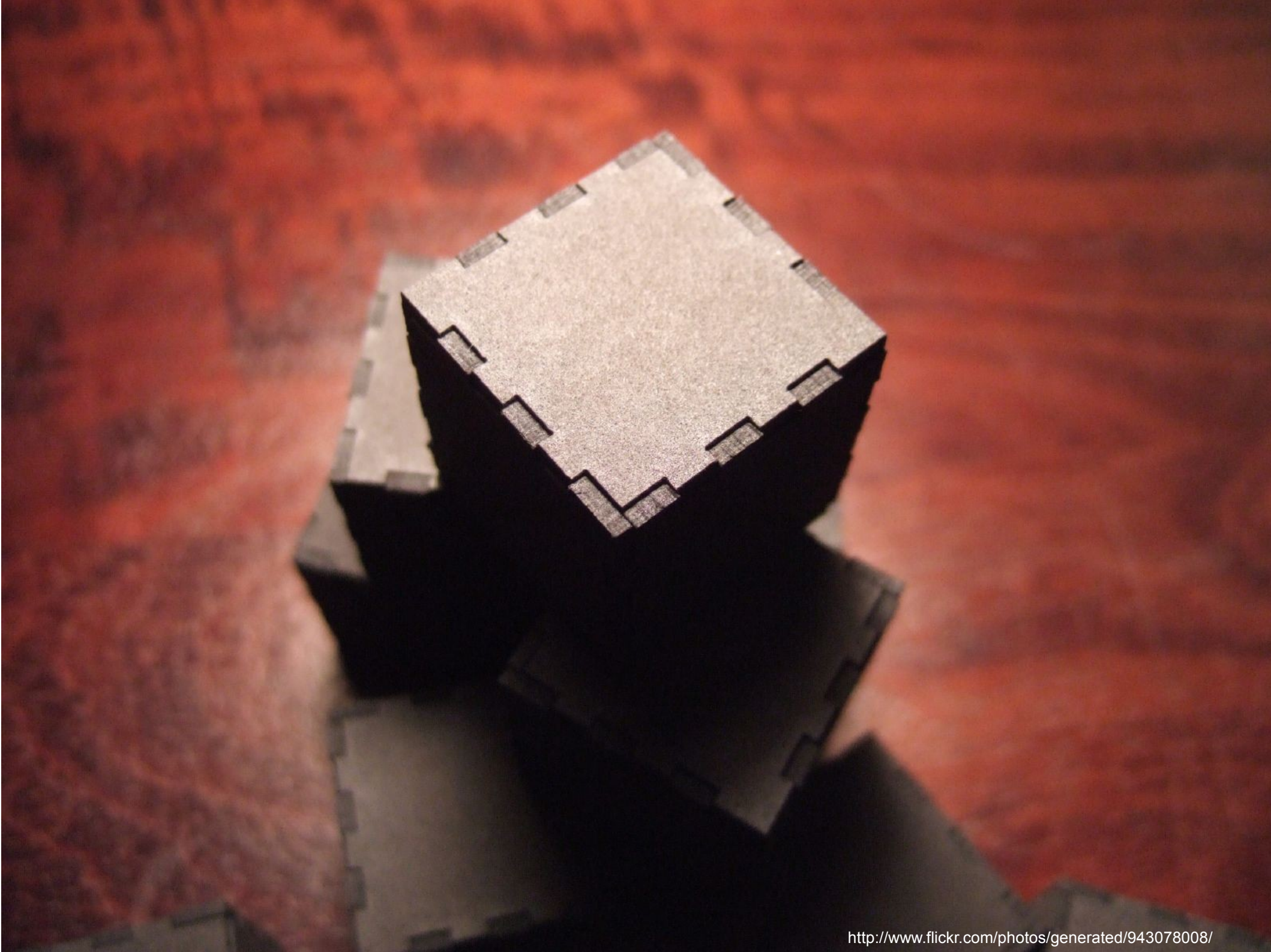


AUC – area under ROC



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

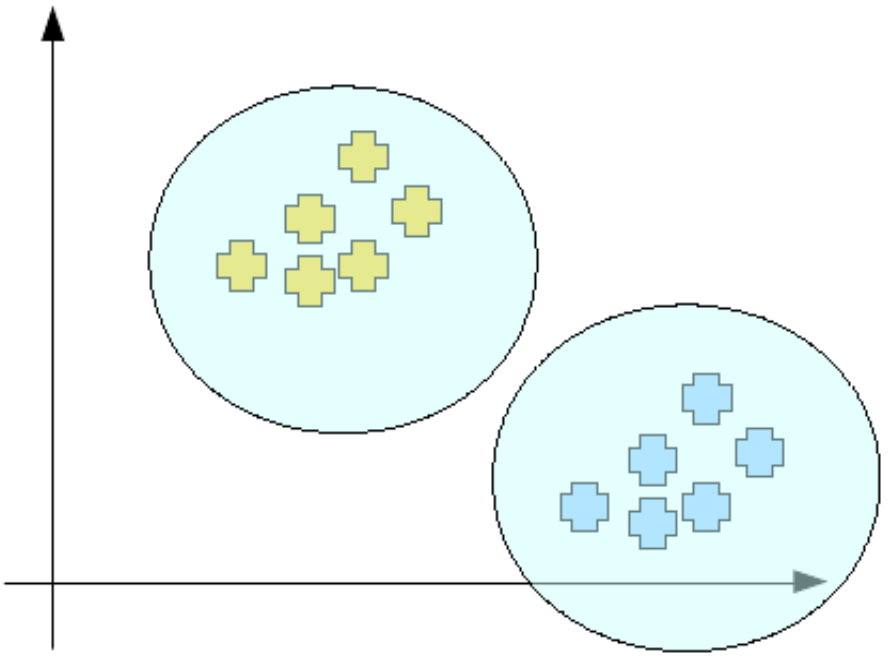
- Given a dataset of texts, identify main topics.

Algorithms: Parallel LDA

- Examples:
  - Dominant topics in set of mails.
  - Identify news message categories.

Discover groups of items

- Group items by similarity.



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

U.K.

- Top Stories
- World
- U.K.
- Business
- Sci/Tech
- Entertainment
- Sports
- Health
- Spotlight
- Most Popular

› **All news**
Headlines
Images

World


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

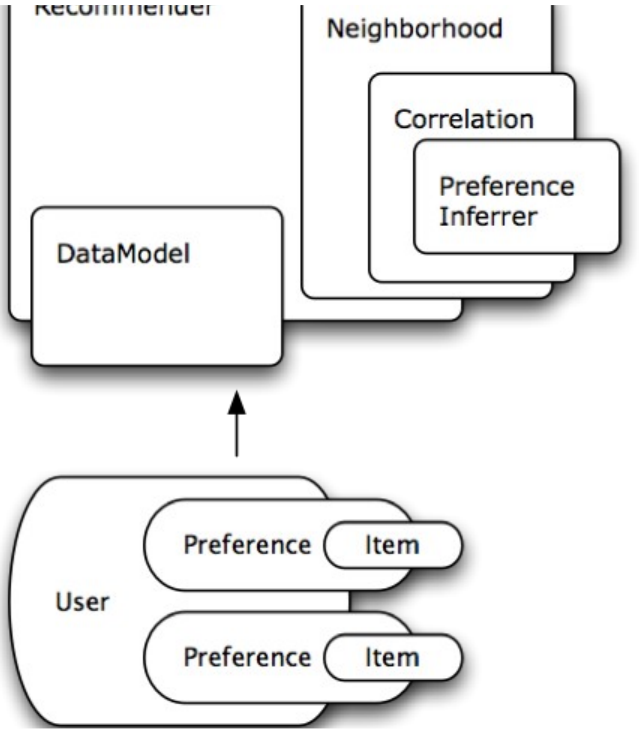
Karadzic snubs his war crimes trial..but it will go ahead without him Mirror.co.uk

guardian.co.uk - New York Times - The Associated Press - Independent

all 1,214 news articles » Email this story

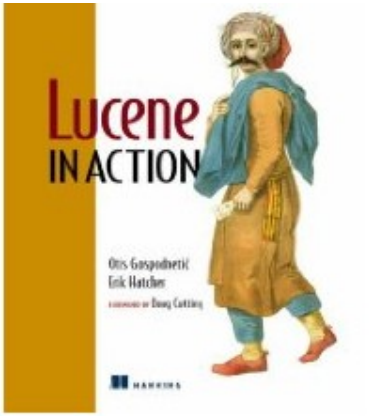
Recommendation mining.

- Collaborative filtering.



Show most relevant ads

Show most relevant ads



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

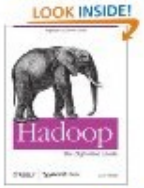
Price For Both: $70.12

Add both to Cart | Add both to Wish List

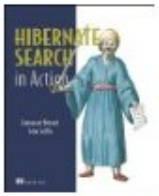
Show availability and shipping details

Customers Who Bought This Item Also Bought

Solr 1.4 Enterprise Search Server by David



Hadoop: The Definitive Guide by Tom White

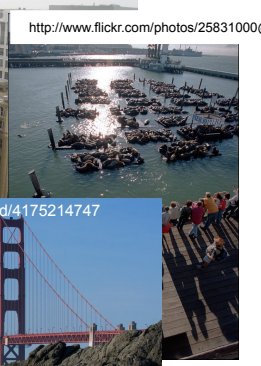


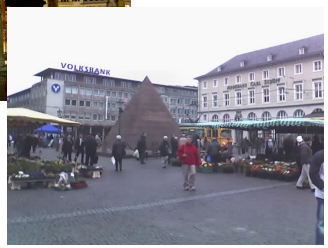
Hibernate Search in Action by Emmanuel

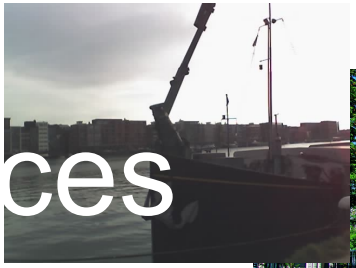
Introduction to Information Retrieval by

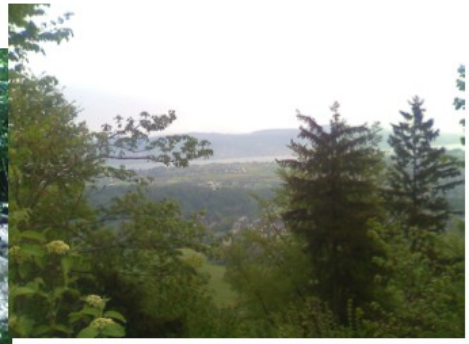
Collective Intelligence in Action by Satnam Alag

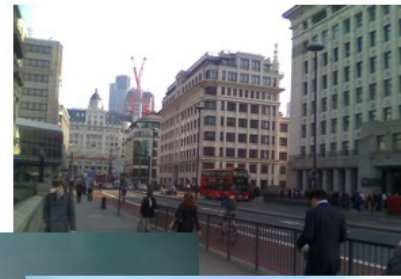
Recommending places

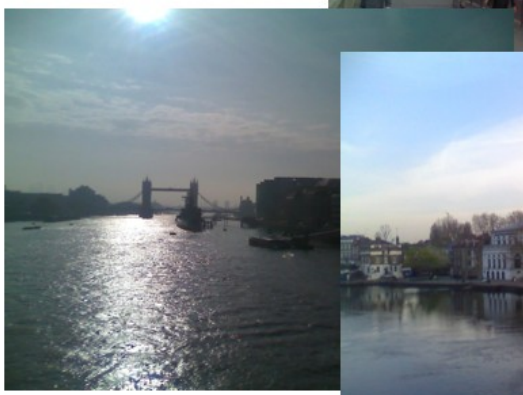


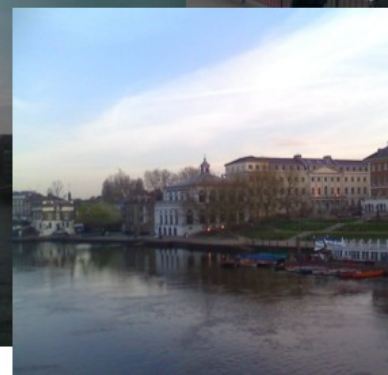












Thanks to Falko Menge for the pictures of Brussels.

Recommending people

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

rypto/3201254932/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/

amazon web services™

Sign in to the AWS Management Console | Create an AWS Account

AWS | Products | Developers | Community | Support | Account

Products & Services

Amazon EC2 Details

- **EC2 Overview**
- FAQs
- Amazon EC2 SLA
- EC2 Instance Types

Amazon Elastic Compute Cloud (Amazon EC2)

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 computing easier for developers.

Amazon EC2's simple web service interface allows you to obtain and 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

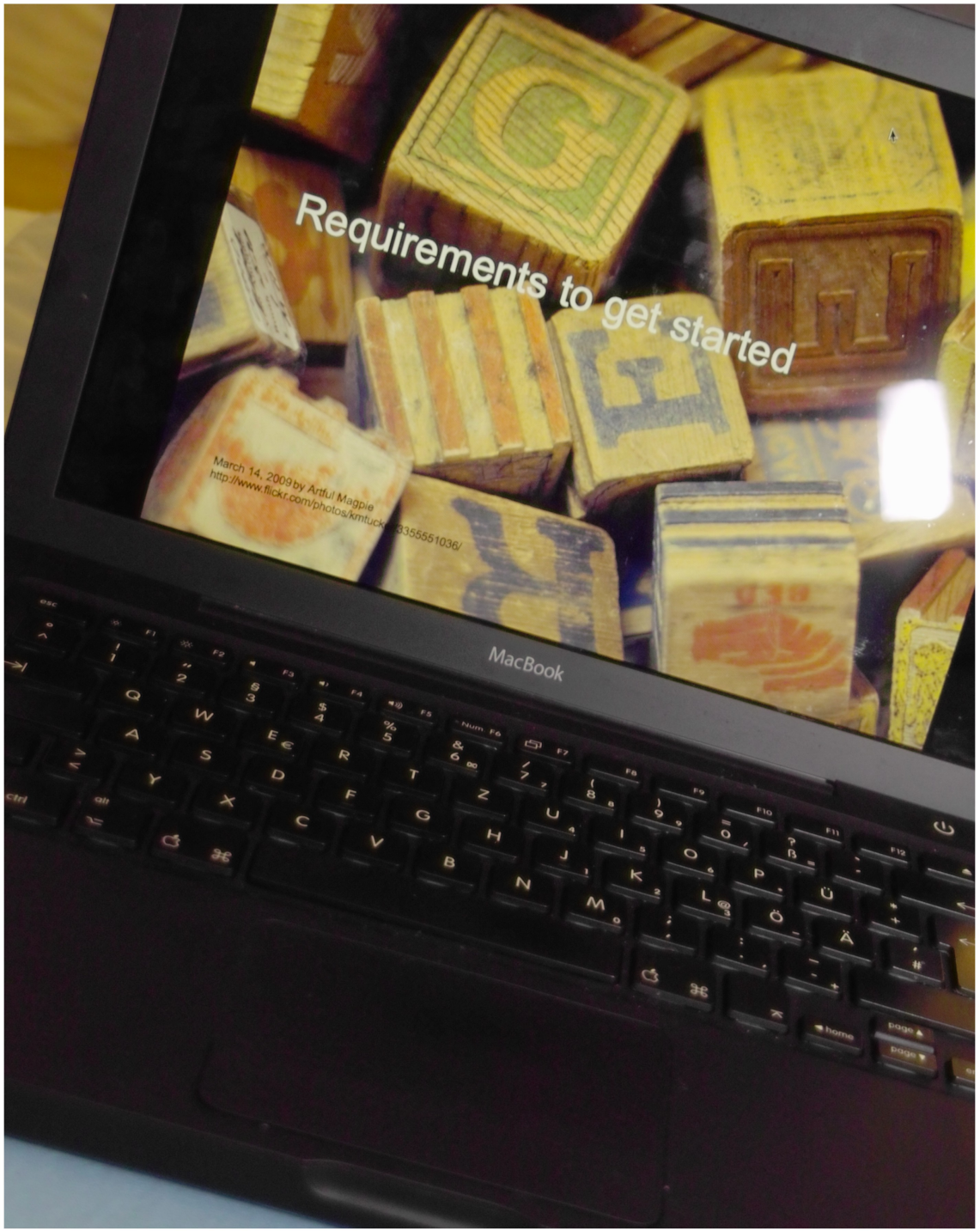
Sign Up For Amazon EC2

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.)


Requirements to get started
March 14, 2009 by Artful Magpie
http://www.flickr.com/photos/kmfuck
3355551036/
MacBook

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




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.


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

I can't recommend this conference enough. Top industry speakers, top developers and fantastic organisation. Mark this event on your sponsoring calendar!" - *Scott Robinson, Senior Marketing Manager, neofonie GmbH*

Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. *Nokia*

Great to have this kind of conferences here in Berlin. Enjoyed to get a good and overview about the various NoSQL options.

Berlin Buzzwords

The conference gave me a good overview on all kinds of scalable open-source projects.

Search/ Store/ Scale

The Buzzwords conference last year put Berlin on the map as Europe's perhaps most important hub for startups and cutting edge web technology today. Already looking forward to the next!
- *Eric Wahlforss, Soundcloud*

June 2011

I enjoyed it very much: Very good location, decent-sized auditoriums, very good wifi, practically all talks were very good: deep expertise and mostly very good presenting skills. I will definitely try to attend again if the event is continued next year. *Nokia*

Berlin Buzzwords 2010 was a great opportunity to showcase our initial release of Lily - a NoSQL content repository based on HBase and SOLR. The event organisation was top-notch: from badges to bag inserts, bannering, food, videotaping - the organisers went out of their way to accommodate both audience, speaker and sponsors in a highly professional way, while still achieving the easy-going, content-above-form atmosphere of a grassroots conference." *Steven Noels - managing partner - Outerthought*

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.