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.

dicode

- “Mastering Data-Intensive Collaboration and Decision Making”

- EU funded research project
  - Number of partners: 8
  - Coordinator: Research Academic Computer Technology Institute (CTI), Greece

neofonie*

Hello Apache Con!

Hello Apache Con!

Lucene

Hello Apache Con!

Apache Solr

Hello Apache Con!

Machine learning background?

Hello Apache Con!

mahout

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?

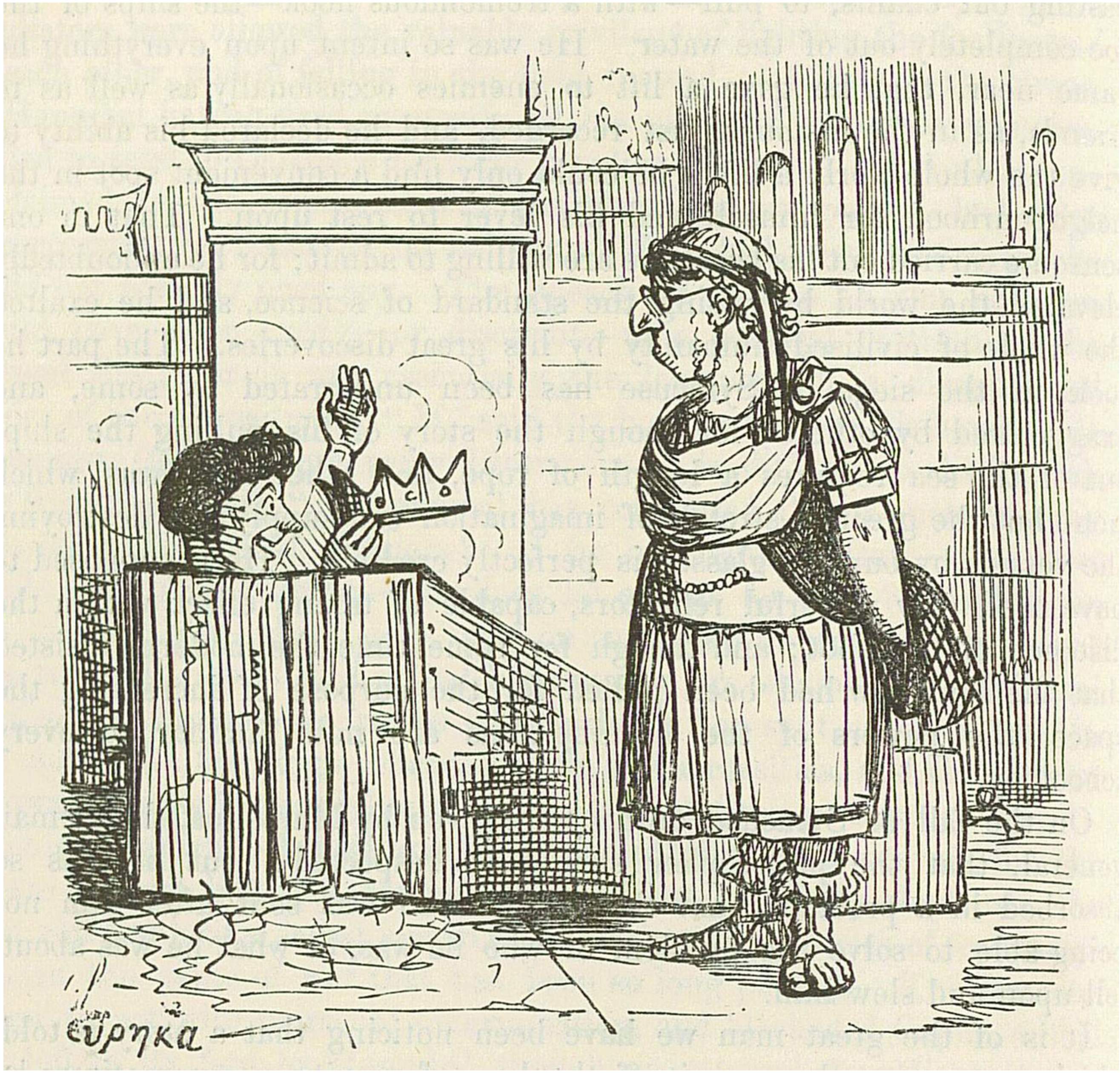


Image by John Leech, from: The Comic History of Rome by Gilbert Abbott A Beckett.
Bradbury, Evans & Co, London, 1850s
Archimedes taking a Warm Bath

Archimedes model of nature

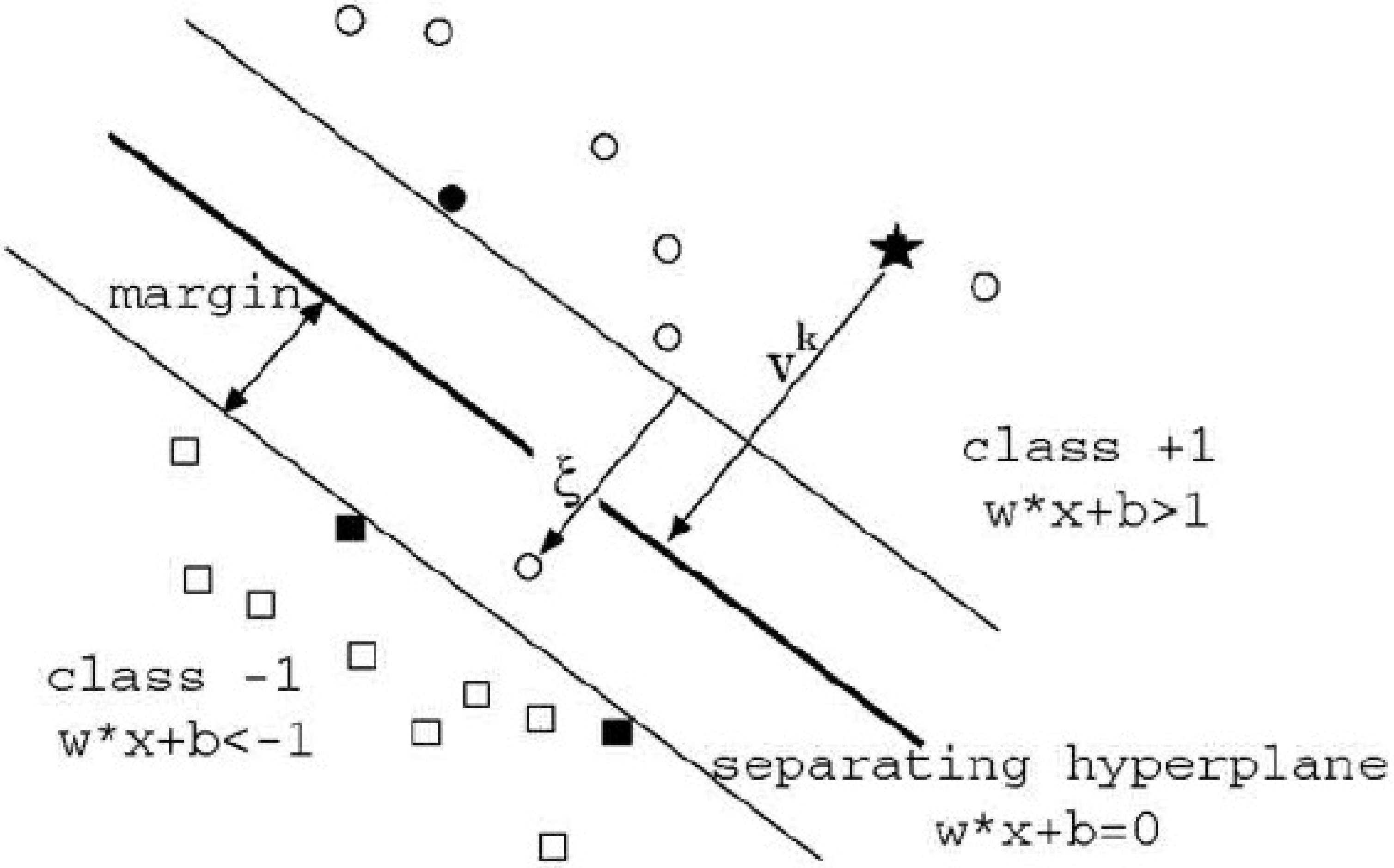
$$\frac{\mathit{Density\ of\ Object}}{\mathit{Density\ of\ Fluid}} = .$$

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

the five people you meet in heaven
Nicholas Sparks A Walk to Remember
Nicholas Sparks The Notebook
Lauren Weisberger The Devil Wears Prada
Martha Moody Best Friends
Maya Angelou The Heart of a Woman
comeback
Christopher Andersen American Evita

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

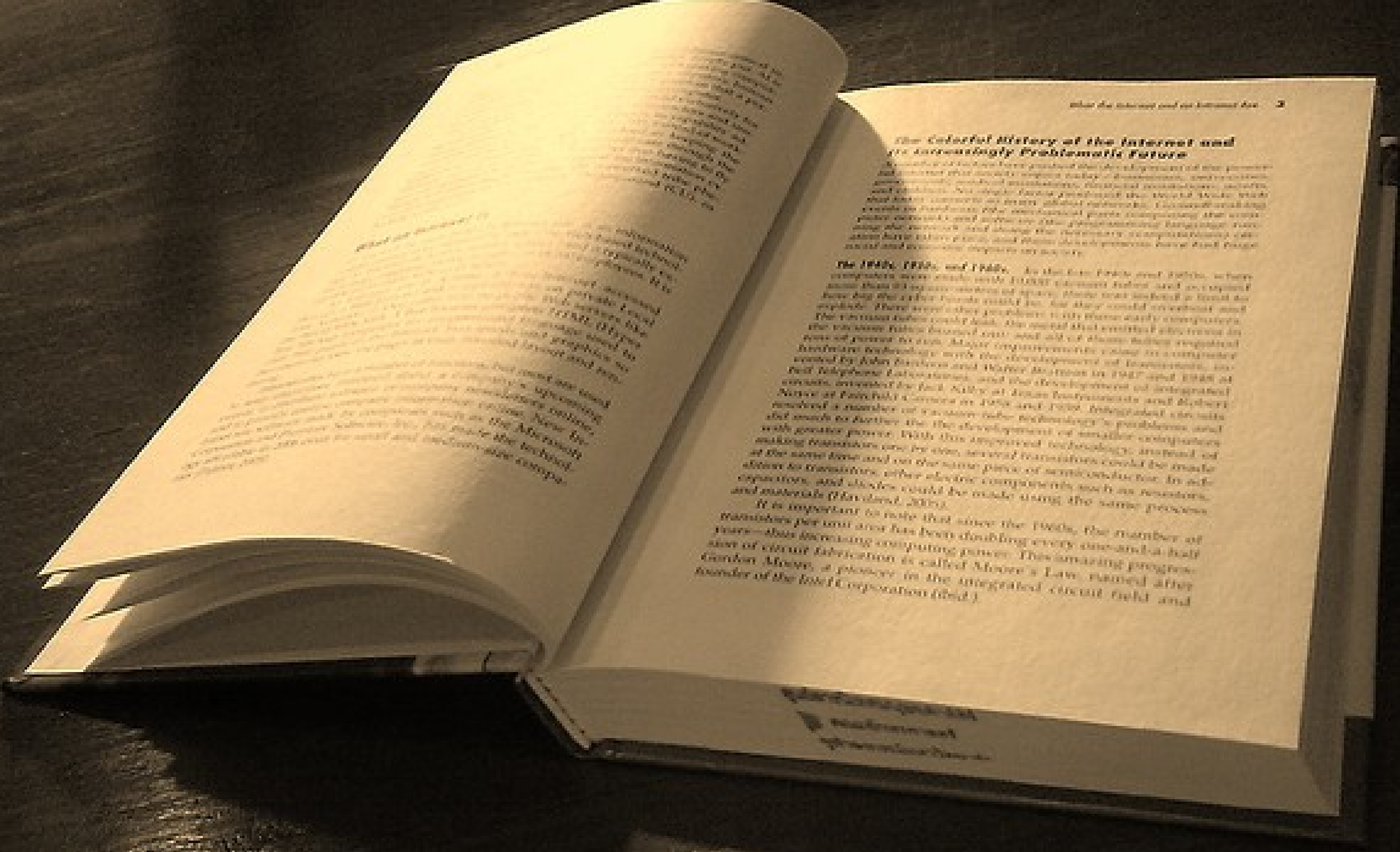
An SVM's model of nature



The challenge

Mission

Provide scalable data mining algorithms.



The Colorful History of the Internet and Its Increasingly Problematic Future

[illegible]

The 1940s, 1950s, and 1960s. In the late 1940s and 1950s, when computers were made with 10,000 vacuum tubes and occupied more than 93 square meters of space, there was indeed a limit to how big the cables could be, for they could overheat and explode. There were other problems with these early computers: the vacuum tubes could leak, the metal that emitted electrons in the vacuum tubes burned out, and all of these tubes required tons of power to run. Major improvements came in computer hardware technology with the development of transistors, invented by John Bardeen and Walter Brattain in 1947 and 1948 at Bell Telephone Laboratories, and the development of integrated circuits, invented by Jack Kilby at Texas Instruments and Robert Noyce at Fairchild Camera in 1958 and 1959. Integrated circuits resolved a number of vacuum-tube technology's problems and did much to further the the development of smaller computers with greater power. With this improved technology, instead of making transistors one by one, several transistors could be made at the same time and on the same piece of semiconductor. In addition to transistors, other electric components such as resistors, capacitors, and diodes could be made using the same process and materials (Haviland, 2005).

It is important to note that since the 1960s, the number of transistors per unit area has been doubling every one-and-a-half years—this increasing computing power. This amazing progression of circuit fabrication is called Moore's Law, named after Gordon Moore, a pioneer in the integrated circuit field and founder of the Intel Corporation (ibid.)

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

HowTo: From data to information.

COMMUNITY NEWS

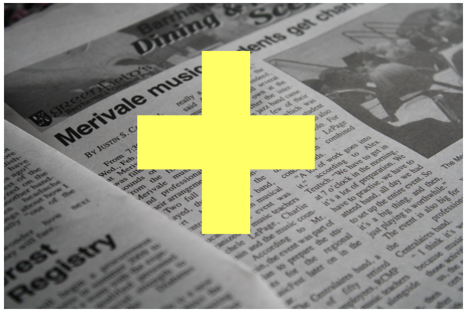
Finishing touches still to come

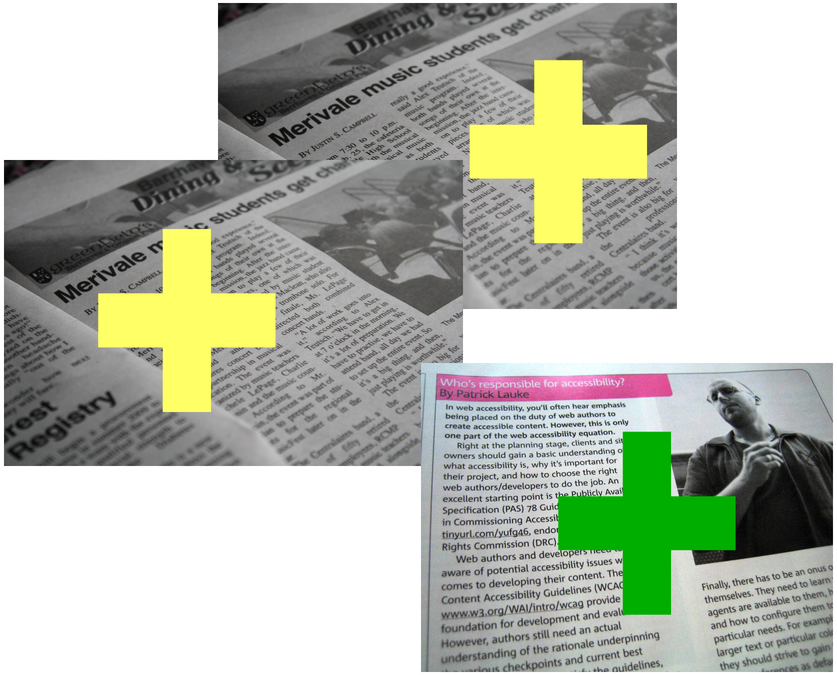
A glimpse of today, yesterday

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/





ZEIT ONLINE | DATENSCHUTZ

Suchen

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

DATENSCHUTZ

Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdaten gegen die Gesundheitskarte: 130 Organisatio ur Demonstration aufgerufen. Sie fürcht Überwachungsstaat.

meier/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 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

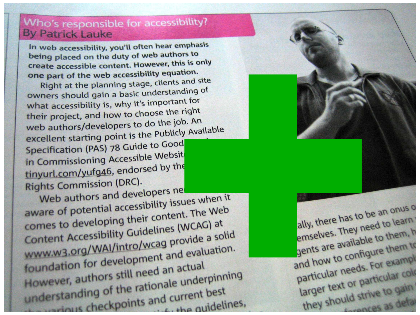
1. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte im Netz
NE Kurier darf ORF-Portal Futurezone kaufen
RECHTE "Wir haben als Kind gelernt, Teilen ist
FACEBOOK "Ein iPad umsonst, ich halte es in
n"
IPAD APPS Mehr Freiheit im App-Store

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

ANZEIGE

experteer.de

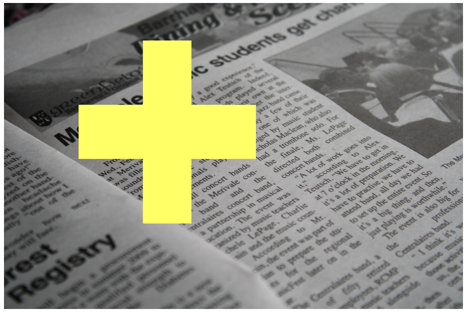

Who's responsible for accessibility?
By Patrick Lauke

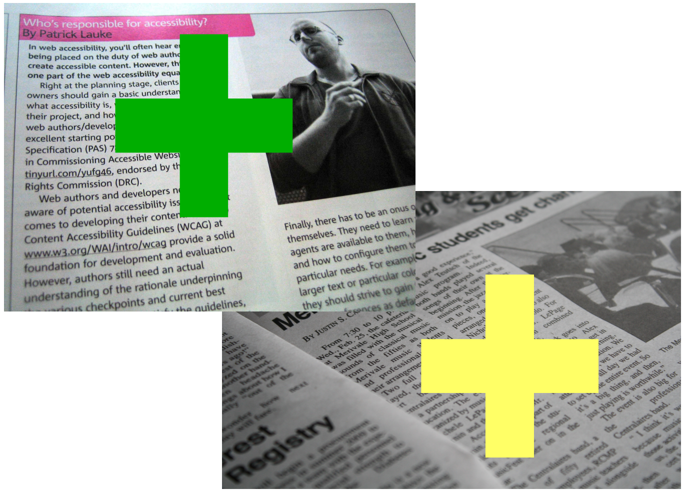
In web accessibility, you'll often hear emphasis being placed on the duty of web authors to create accessible content. However, this is only one part of the web accessibility equation.

Right at the planning stage, clients and site owners should gain a basic understanding of what accessibility is, why it's important for their project, and how to choose the right web authors/developers to do the job. An excellent starting point is the Publicly Available Specification (PAS) 78 Guide to Good in Commissioning Accessible Websit tinyurl.com/yufg46, endorsed by the Rights Commission (DRC).

Web authors and developers ne aware of potential accessibility issues when it comes to developing their content. The Web Content Accessibility Guidelines (WCAG) at www.w3.org/WAI/intro/wcag provide a solid foundation for development and evaluation. However, authors still need an actual understanding of the rationale underpinning the various checkpoints and current best the guidelines,

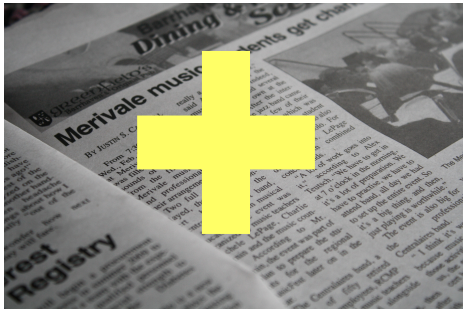
ally, there has to be an onus o emselves. They need to learn gents are available to them, h and how to configure them to particular needs. For exampl larger text or particular col they should strive to gain ferences as defa

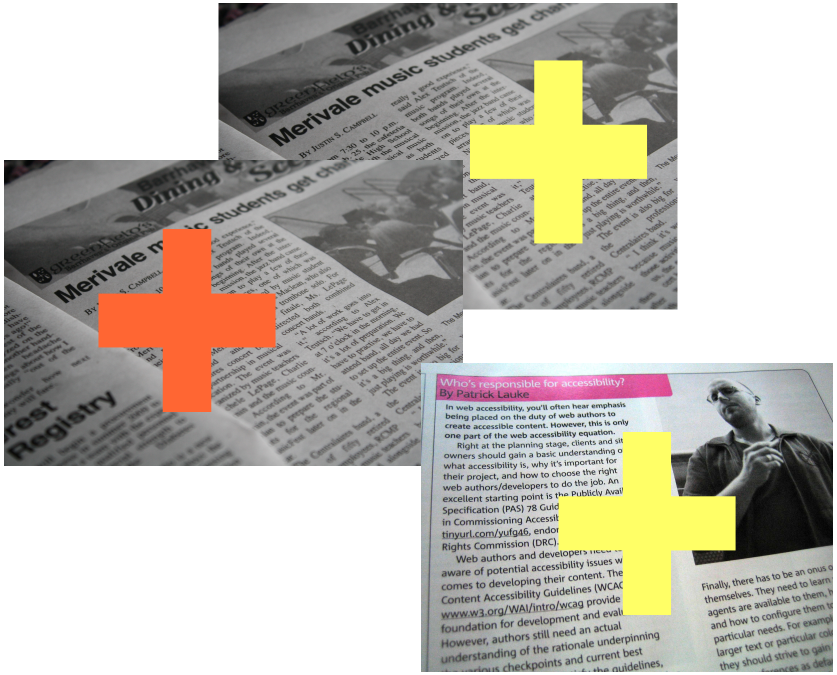


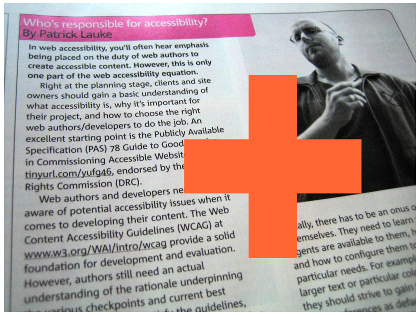


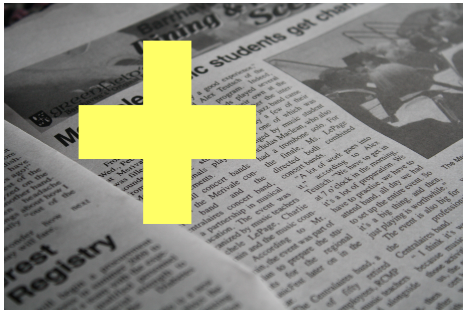
http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

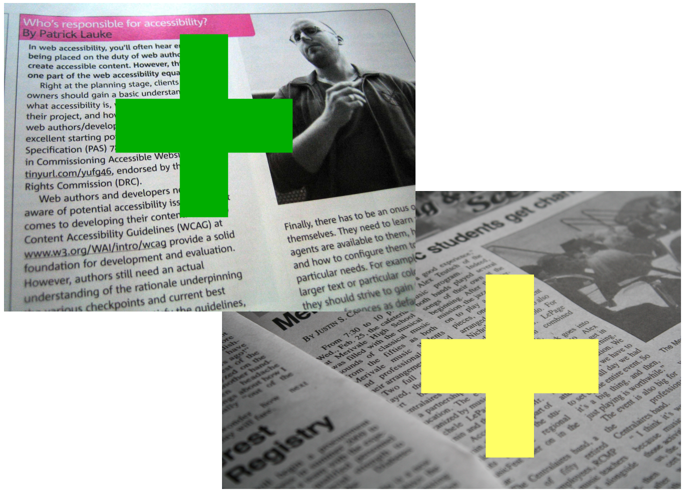
http://www.flickr.com/photos/redux/409356158/




Merivale music students get chance
Dining & Scene
Who's responsible for accessibility?
By Patrick Lauke







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

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

ZEIT ONLINE | DATENSCHUTZ

Suchen

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

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 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

ANZEIGE

Who's responsible for accessibility?
By Patrick Lauke

In web accessibility, you'll often hear emphasis being placed on the duty of web authors to create accessible content. However, this is only one part of the web accessibility equation.

Right at the planning stage, clients and site owners should gain a basic understanding of what accessibility is, why it's important for their project, and how to choose the right web authors/developers to do the job. An excellent starting point is the Publicly Available Specification (PAS) 78 Guide to Good Practice in Commissioning Accessible Websites, at tinyurl.com/yufg46, endorsed by the Disability Rights Commission (DRC).

Web authors and developers need to be aware of potential accessibility issues when it comes to developing their content. The Web Content Accessibility Guidelines (WCAG) at www.w3.org/WAI/intro/wcag provide a solid foundation for development and evaluation. However, authors still need an actual understanding of the rationale underpinning the various checkpoints and current best

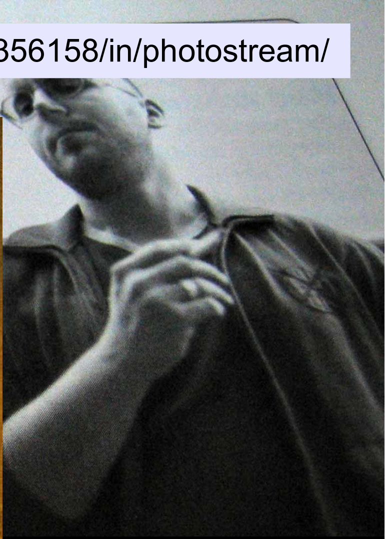
Finally, there has to be an onus o themselves. They need to learn agents are available to them, h and how to configure them to particular needs. For exampl larger text or particular col they should strive to gain

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

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

Dining & Scene
Merivale music students get chance
Who's responsible for accessibility?
By Patrick Lauke
ZEIT ONLINE
DATENSCHUTZ
Tausende demonstrieren für Bürgerrechte im Netz
Freiheit statt Angst

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 Im

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 Ur
3. GEDEN Ressen
4. BUNDESLIGA Der BVB und die Freiburger siegen
5. DAAD Das Ende der Ära Bode

New China
Tim Griffith/PTW Architects
The National Aquatics Center lies on the

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/

ZEIT ONLINE | DATENSCHUTZ

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

DATENSCHUTZ

Tausende demonstrieren für Bürgerrechte im Netz

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

In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz

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

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

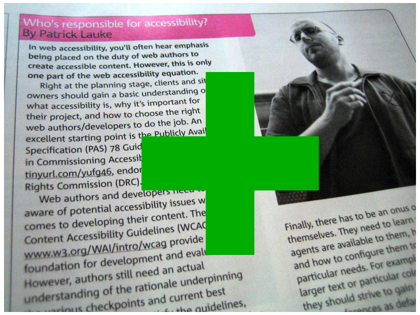
1. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte im Netz
2. FUTUREZONE Kurier darf ORF-Portal Futurezone kaufen
3. URHEBERRECHTE "Wir haben als Kind gelernt, Teilen ist gut"
4. SPAM AUF FACEBOOK "Ein iPad umsonst, ich halte es in den Händen"
5. IPHONE & IPAD APPS Mehr Freiheit im App-Store

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

ANZEIGE

experteer.de



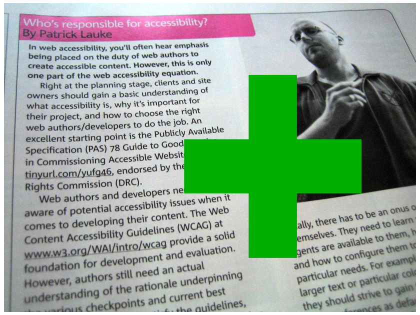
Who's responsible for accessibility?
By Patrick Lauke

In web accessibility, you'll often hear emphasis being placed on the duty of web authors to create accessible content. However, this is only one part of the web accessibility equation.

Right at the planning stage, clients and site owners should gain a basic understanding of what accessibility is, why it's important for their project, and how to choose the right web authors/developers to do the job. An excellent starting point is the Publicly Available Specification (PAS) 78 Guide to Good in Commissioning Accessible Websit tinyurl.com/yufg46, endorsed by the Rights Commission (DRC).

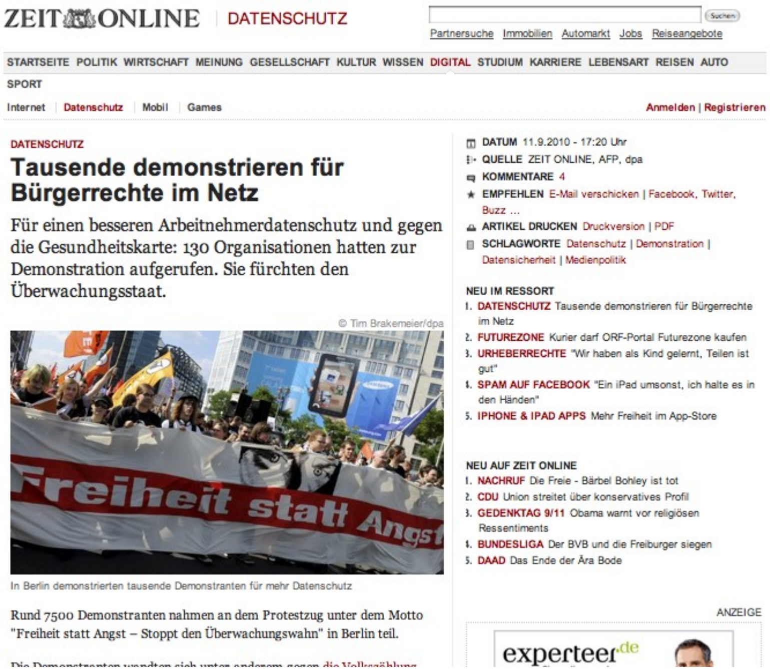
Web authors and developers ne aware of potential accessibility issues when it comes to developing their content. The Web Content Accessibility Guidelines (WCAG) at www.w3.org/WAI/intro/wcag provide a solid foundation for development and evaluation. However, authors still need an actual understanding of the rationale underpinning the various checkpoints and current best

Finally, there has to be an onus o themselves. They need to learn agents are available to them, h and how to configure them to particular needs. For exampl larger text or particular col they should strive to gain



From data to information.

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



ZEIT ONLINE | DATENSCHUTZ

Suchen

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

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

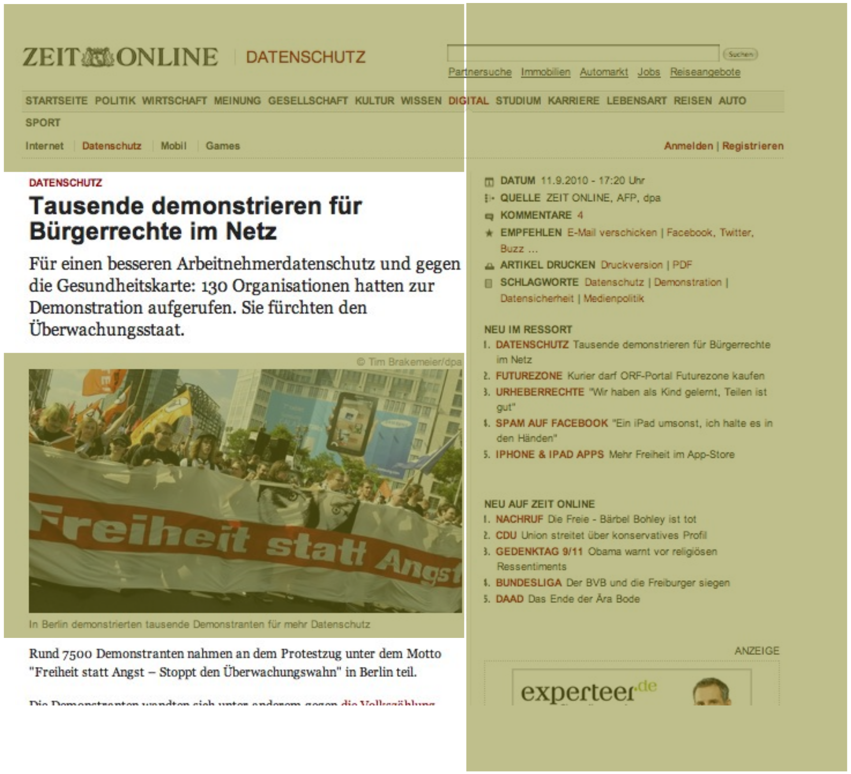
SCHLAGWORTE Datenschutz | Demonstration | Datensicherheit | Medienpolitik

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



ZEIT ONLINE | DATENSCHUTZ

Suchen

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

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

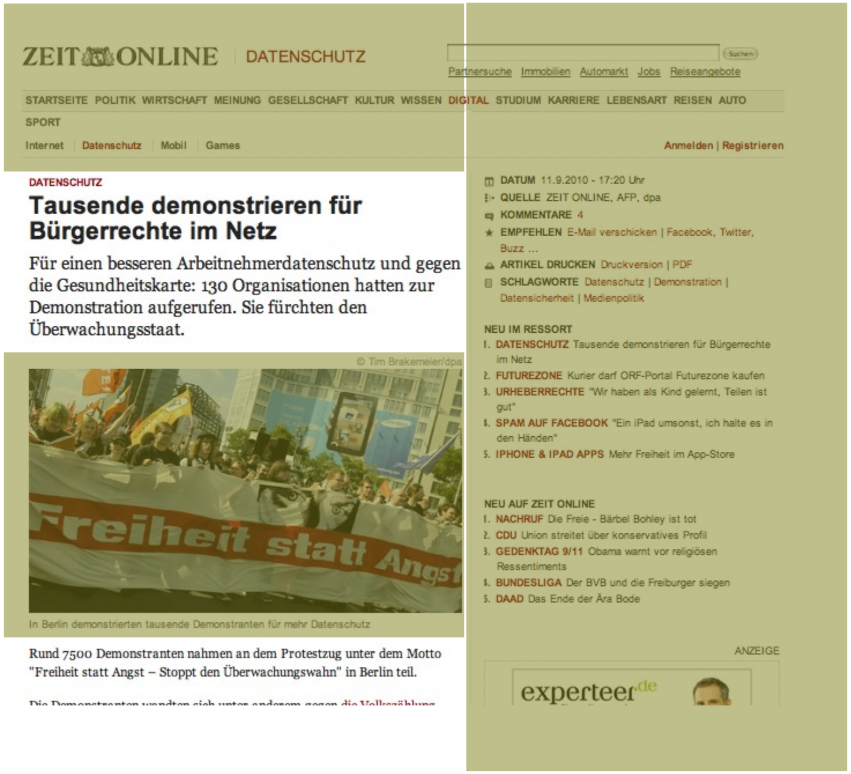
NEU IM RESSORT

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.



ZEIT ONLINE | DATENSCHUTZ

Suchen

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

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

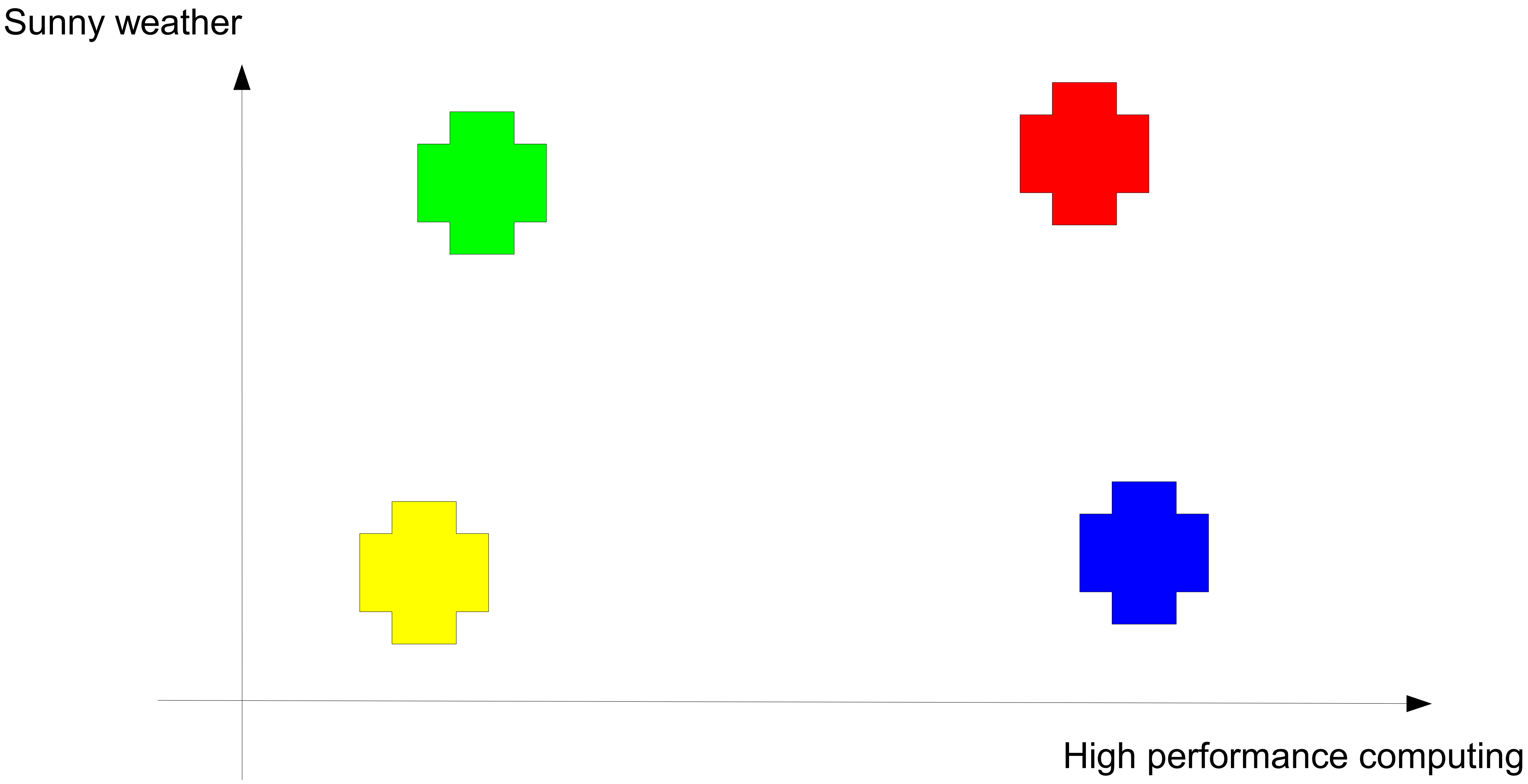
NEU AUF ZEIT ONLINE

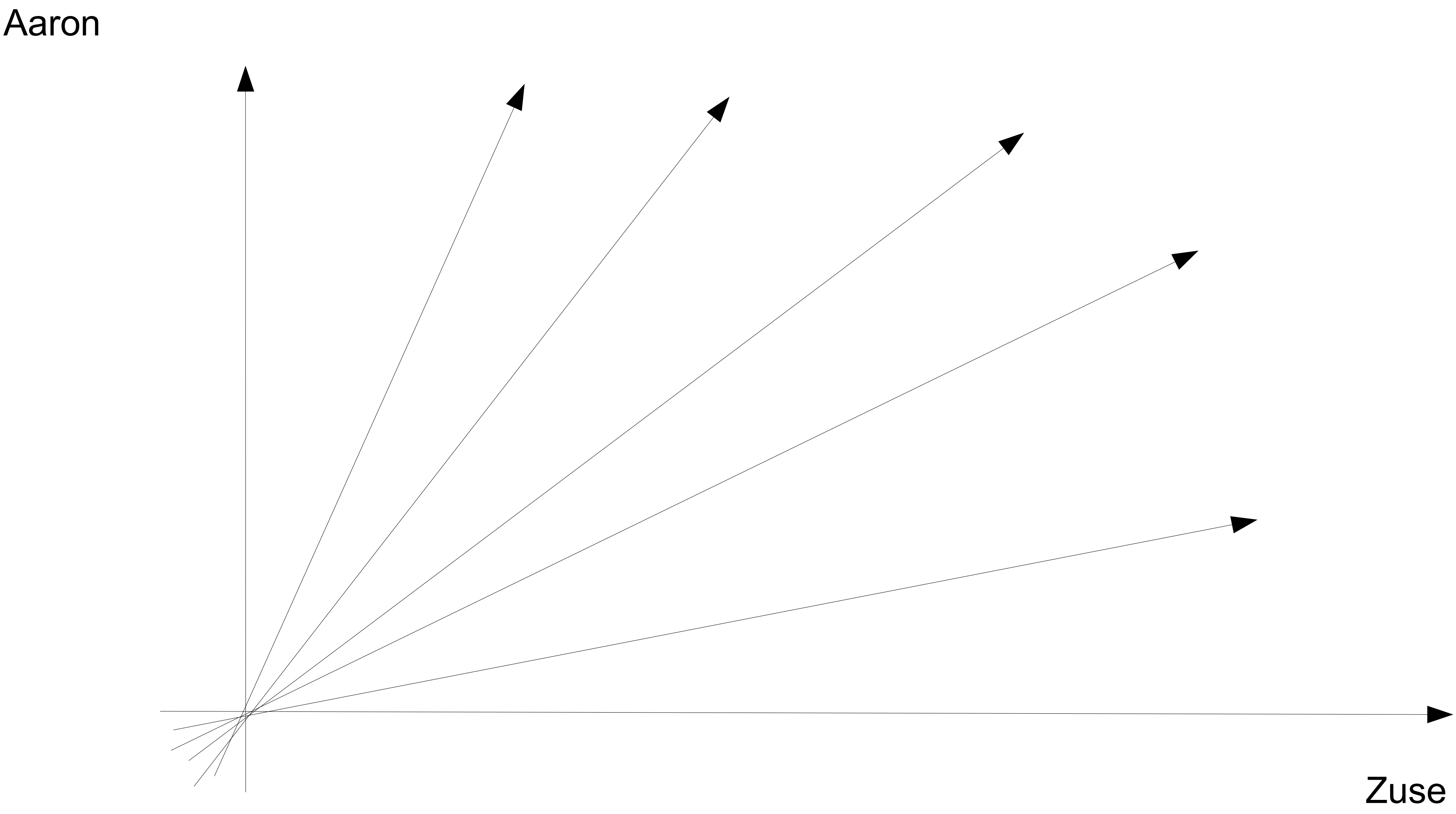
1. NACHRUF Die Freie - Bärbel Bohley ist tot
2. CDU Union streitet über konservatives Profil
3. GEDENKTAG 9/11 Obama warnt vor religiösen Ressentiments
4. BUNDESLIGA Der BVB und die Freiburger siegen
5. DAAD Das Ende der Ära Bode

- Remove noise.
- Convert text to vectors.

From texts to vectors

If we looked at two words only:




Aaron
Zuse

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.

$$b_{i,j} = \begin{Bmatrix} 1 \forall x_i \in d_j \\ 0\,else \end{Bmatrix}$$

- Problem:
  - Number of word occurrences not accounted for.

Term Frequency

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Entry in vector equal to the words frequency.

$$b_{i,j} = n_{i,j}$$

- Problem:
  - Common words dominate vectors.

TF with stop wording

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the words frequency.

$$b_{i,j} = n_{i,j}$$

- Problem:
  - Common and uncommon words with same weight.

TF- IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.

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

- Problem:
  - Long texts get larger values.

Normalized TF- IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.
- Normalize vectors.

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

- Problem:
  - Additional domain knowledge ignored.

Reality

- There are a few more words in news.
- Use all relevant features/ signals available.
  - Words.
  - Header fields.
  - Characteristics of publishing url.
  - …
- Usually pipeline of feature extractors.

From data to information.

- ✓ Collect data and define your learning problem.
- ✓ 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.

ZEIT ONLINE | DATENSCHUTZ

Partnersuche Immobilien Automarkt Jobs Reiseangebote

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

Internet | Datenschutz | Mobil | Games

Anmelden | Registrieren

DATENSCHUTZ

Tausende demonstr Bürgerrechte im Ne

Für einen besseren Arbeitnehm z und gegen die Gesundheitskarte: 130 Org atten zur Demonstration aufgerufen. Sie Überwachungsstaat.

In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz

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

DATUM 11.9.2010 - 17:20 Uhr

QUELLE ZEIT ONLINE, AFP, dpa

KOMMENTARE 4

EMPFEHLEN E-Mail verschicken | Buzz ...

ARTIKEL DRUCKEN Druckversion |

SCHLAGWORTE Datenschutz | Der Datensicherheit | Medienpolitik

NEU IM RESSORT

ENSCHUTZ Tausende demonstri Netz

TUREZONE Kurier darf ORF-Porta

EBERRECHTE "Wir haben als Ki

AM AUF FACEBOOK "Ein iPad un Händen"

ONE & IPAD APPS Mehr Freiheit

NEU AUF ZEIT ONLINE

1. NACHRUF Die Freie - Bärbel Bohley i
2. CDU Union streitet über konservative
3. GEDENKTAG 9/11 Obama warnt vor Ressentiments
4. BUNDESLIGA Der BVB und die Freit
5. DAAD Das Ende der Ära Bode

experteer.de

Who's responsible for accessibility?
By Patrick Lauke

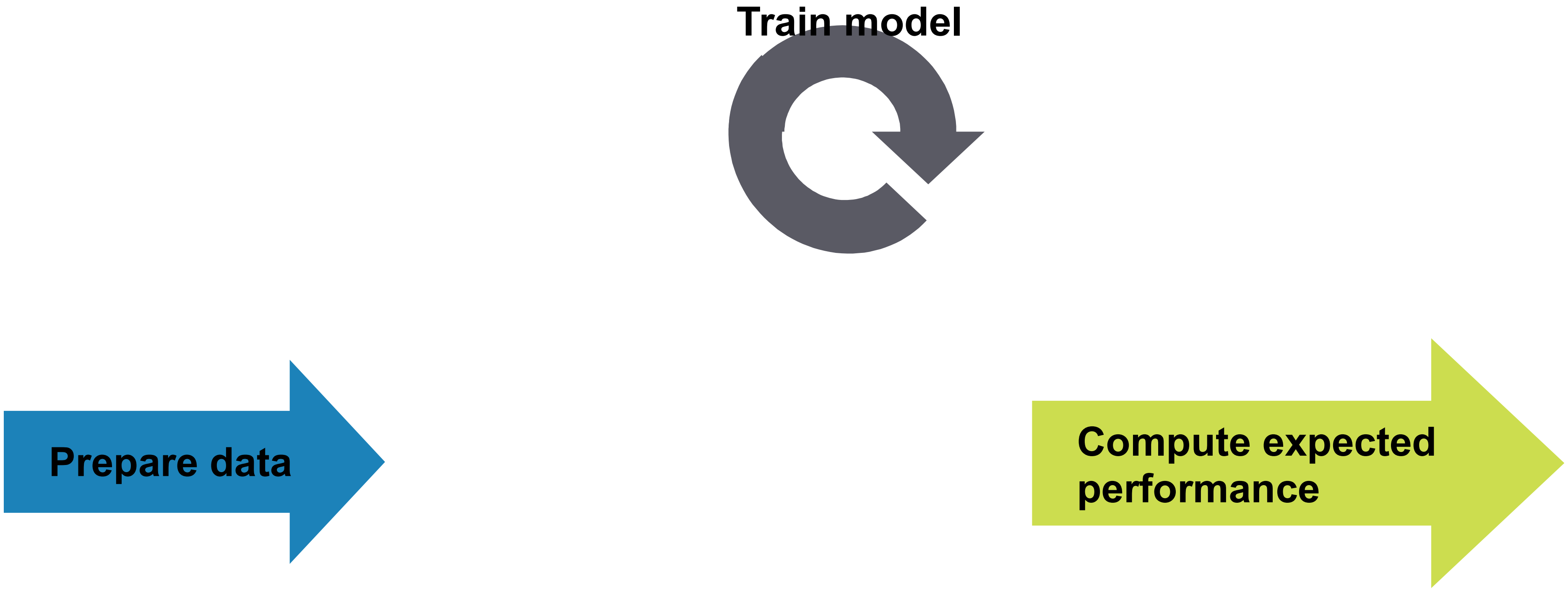
In web accessibility, you'll often hear emphasis being placed on the duty of web authors to create accessible content. However, this is only one part of the web accessibility equation.

Right at the planning stage, clients and site owners should gain a basic understanding of what accessibility is, why it's important for their project, and how to choose the right web authors/developers to do the job. An excellent starting point is the Publicly Available Specification (PAS) 78 Guide to Good Practice in Commissioning Accessible Websites, tinyurl.com/yufg46, endorsed by the Di Rights Commission (DRC).

Web authors and developers need t aware of potential accessibility issues when it comes to developing their content. The Web Content Accessibility Guidelines (WCAG) at www.w3.org/WAI/intro/wcag provide a solid foundation for development and evaluation. However, authors still need an actual understanding of the rationale underpinning the various checkpoints and current best the guidelines,

Goals

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

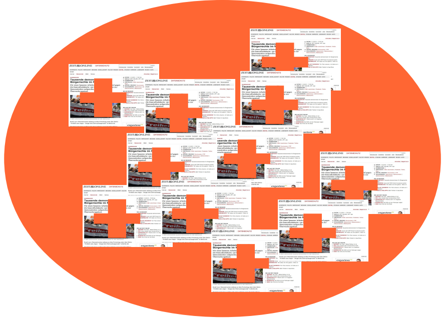


Prepare data

Train model

Compute expected performance

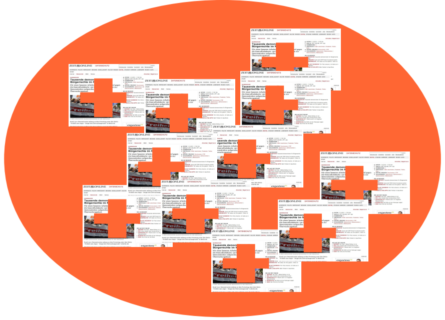
Performance



- Use same data for training and testing.

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

Performance

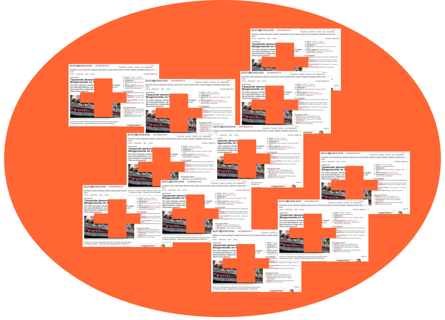


- Use same data for training and testing.

DON'T

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

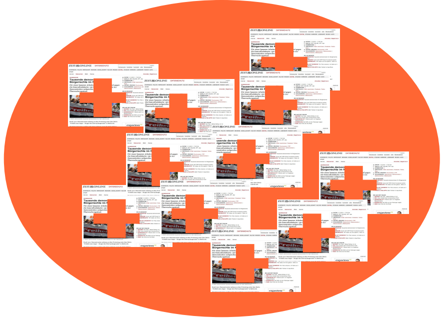
Performance



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

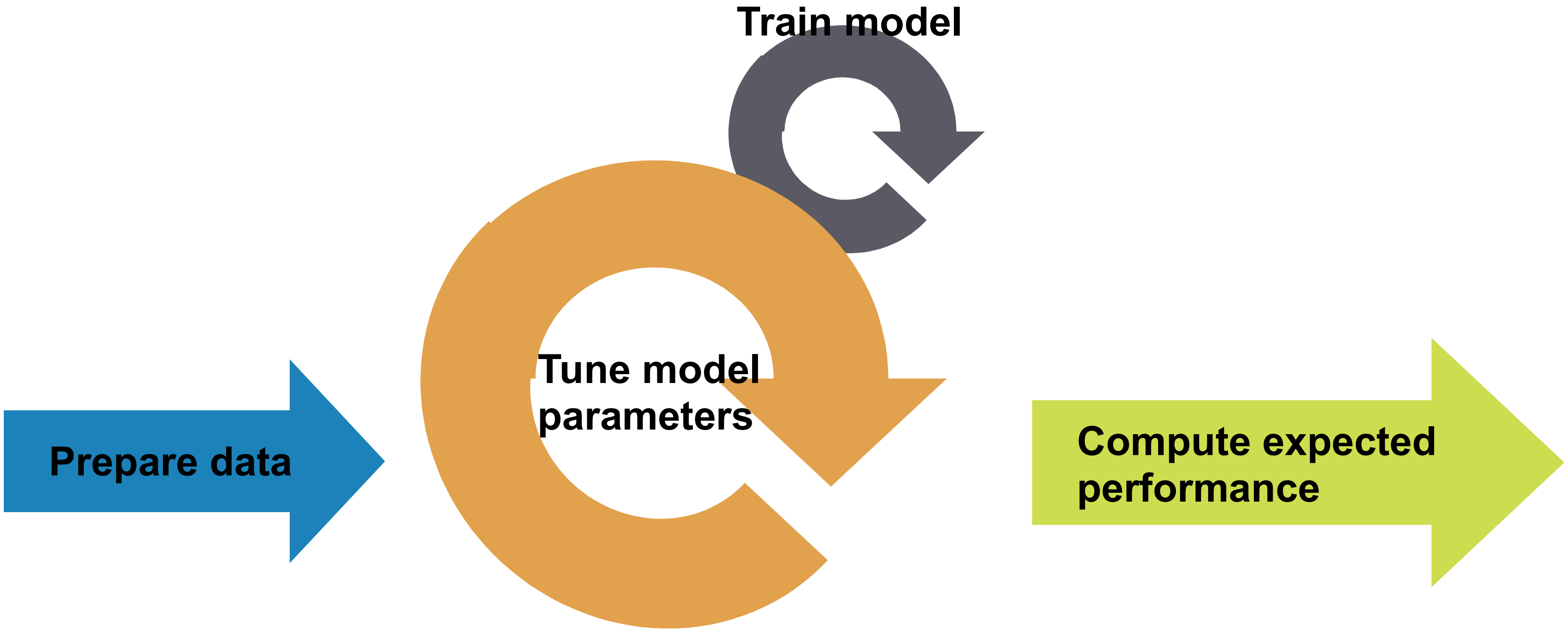
- Problems:
  - Pessimistic predictor: Not all data used for training.
  - Result may depend on which data was set aside.

Performance



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

- Problem:
    - Still a pessimistic predictor.



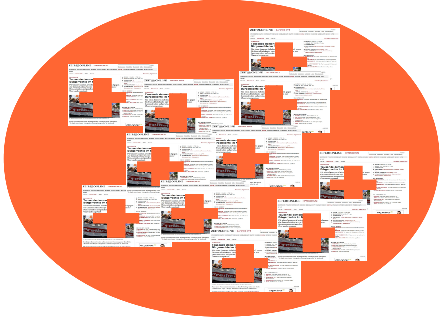
Train model

Prepare data

Tune model parameters

Compute expected performance

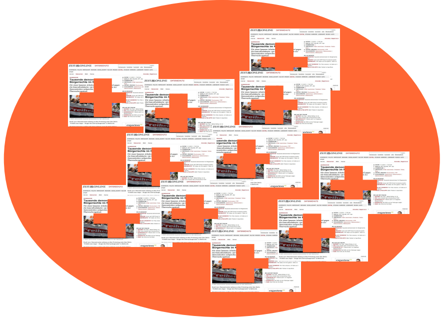
Performance



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

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

Performance

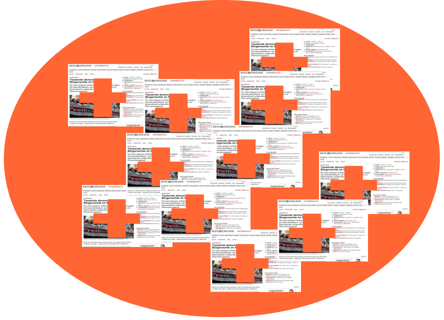


- Use just a fraction for trai g.
- Set some data and testing.

DON'T

- Problems:
  - Highly optimis tic.
  - 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.

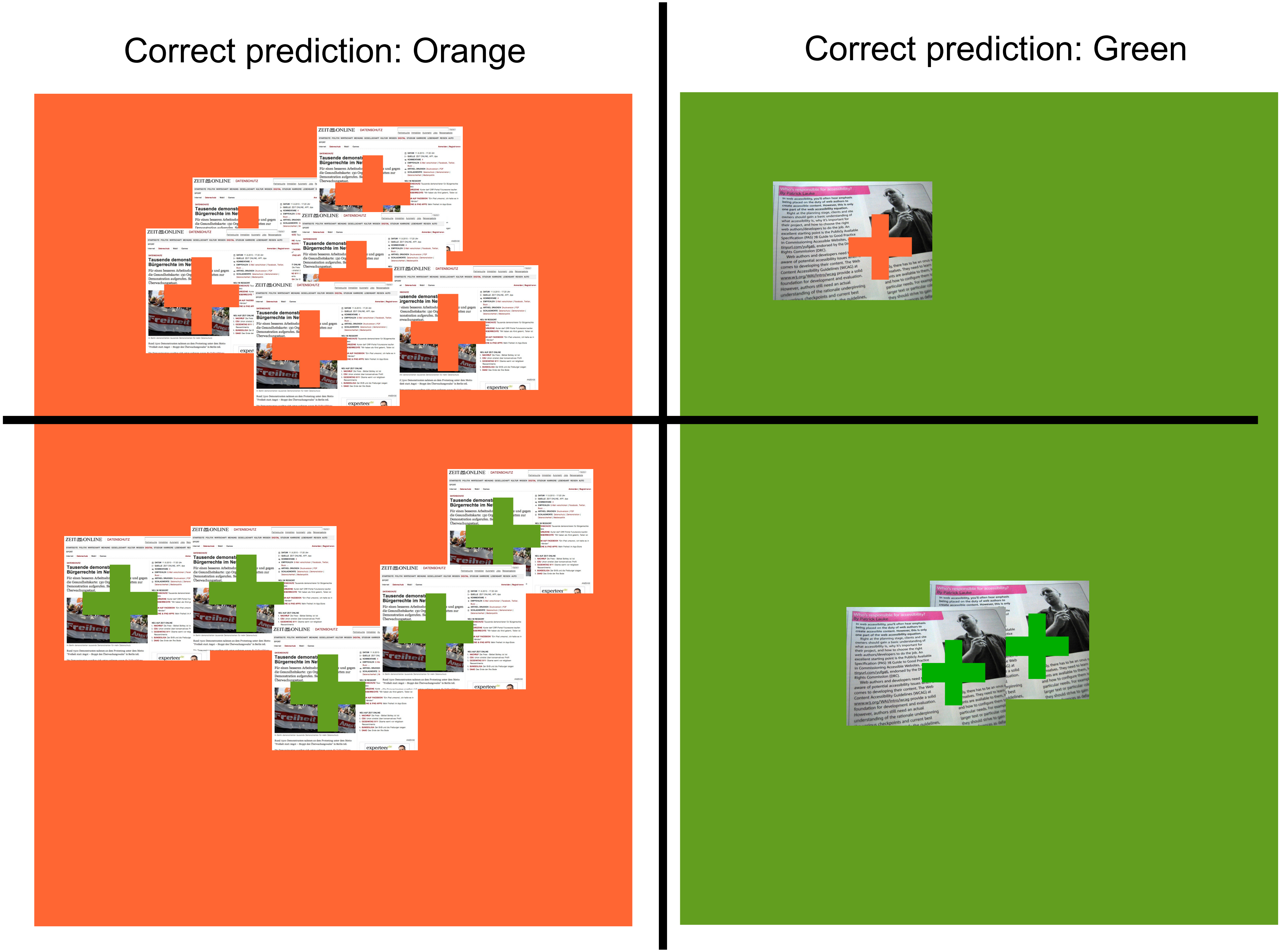
Performance Measures

Correct prediction: Orange

Correct prediction: Green

Model prediction: Orange

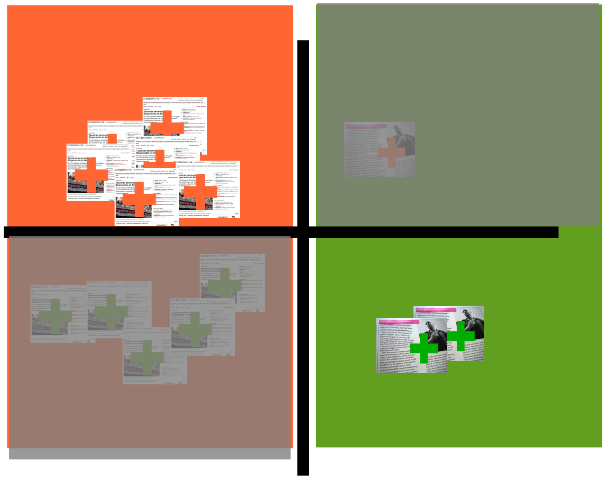
Model prediction: Green



Accuracy

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

- Problems:
  - What if class distribution is skewed?

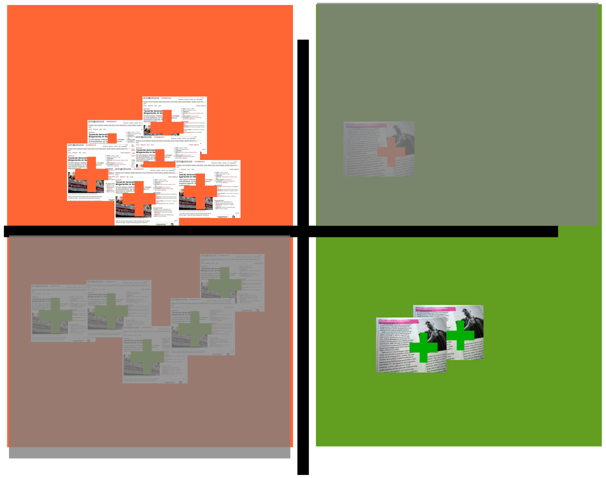


Precision/ Recall

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

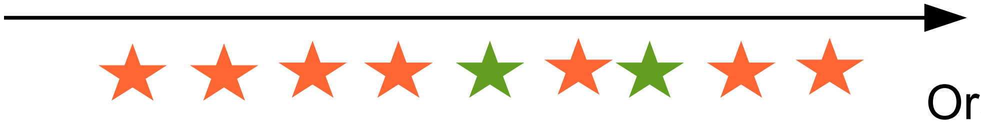
$$Recall = \frac{true\ positive}{true\ positive + false\ negative}$$

- Problem:
  - Depends on decision threshold.



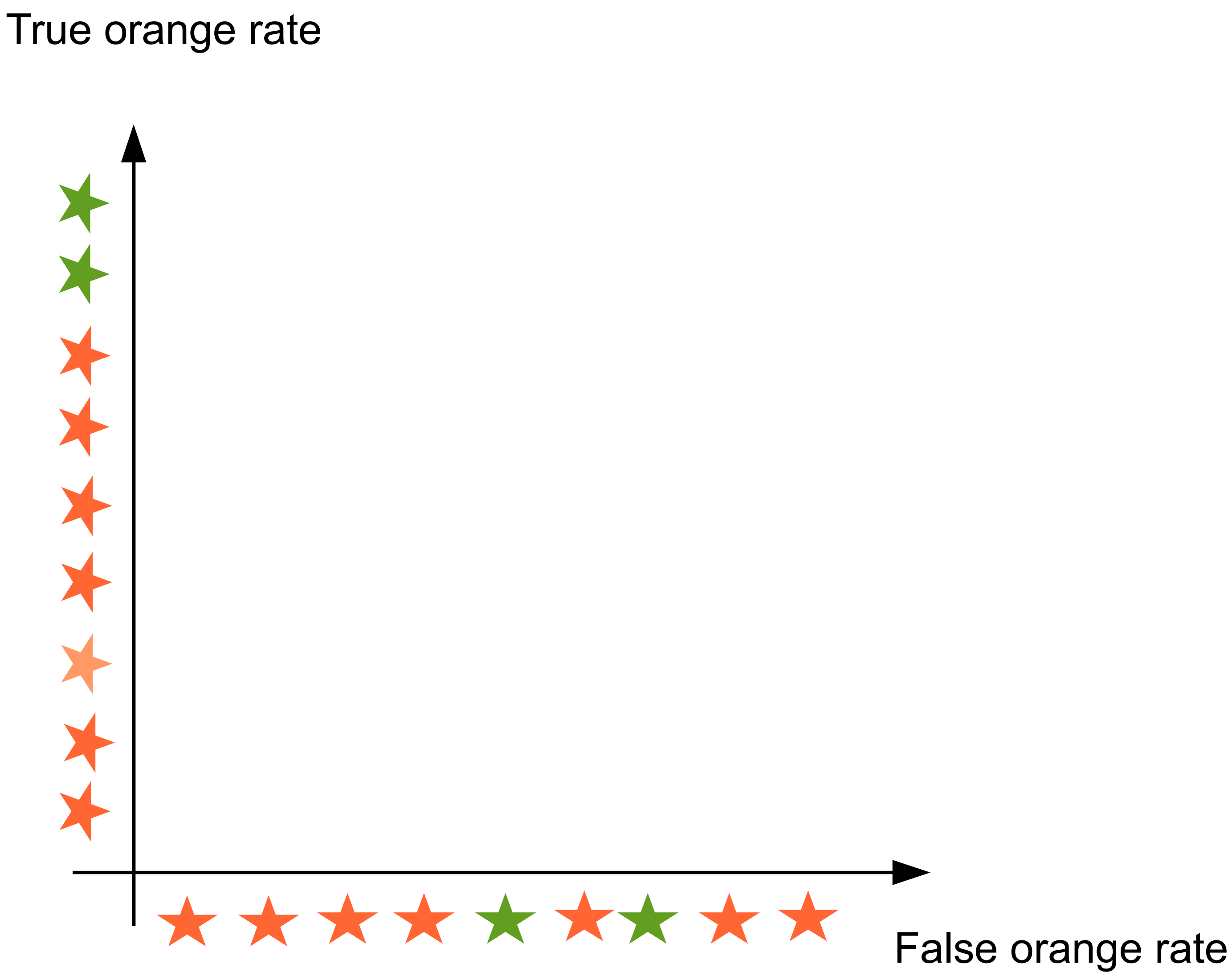
ROC Curves

ROC Curves

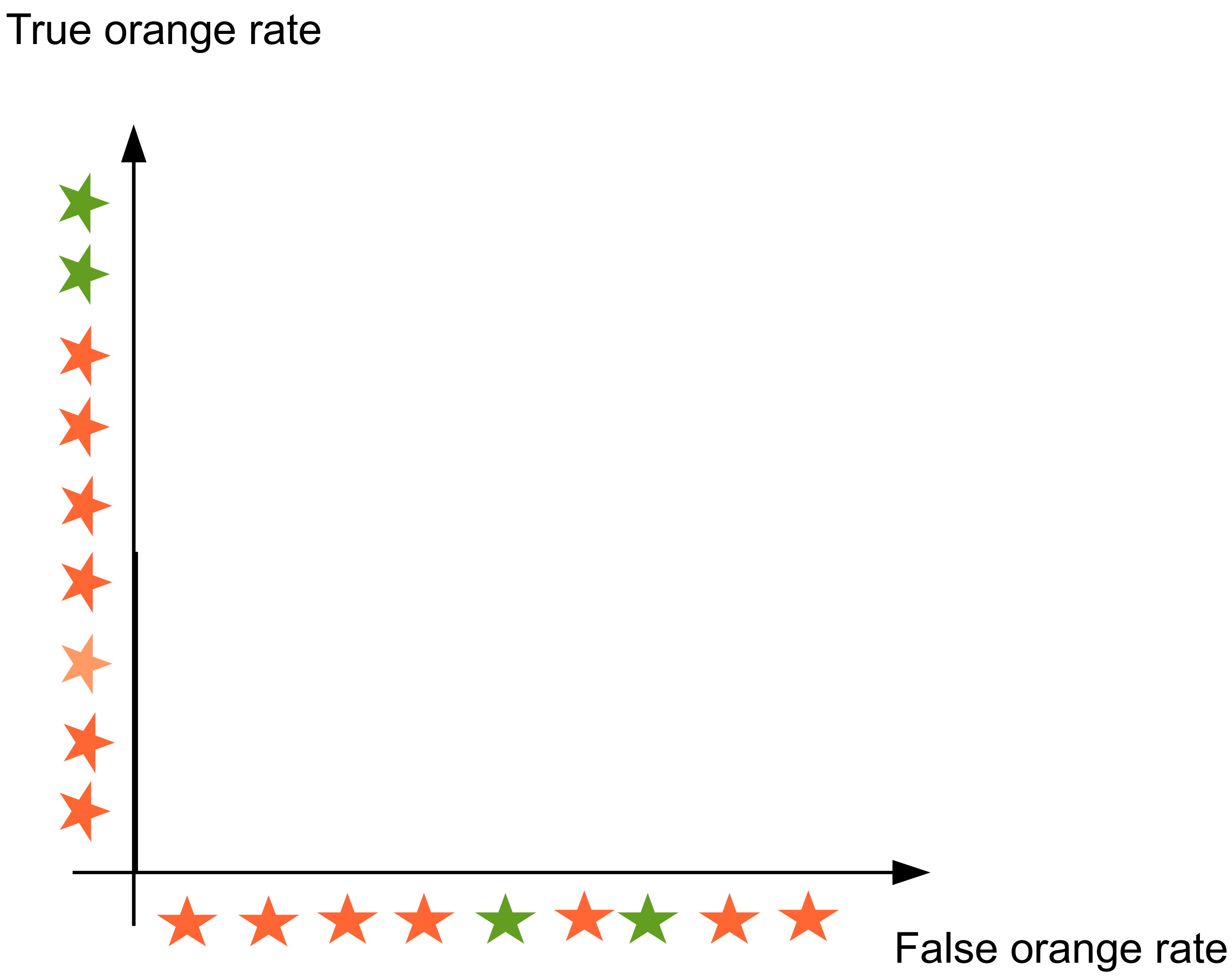


Orange rate

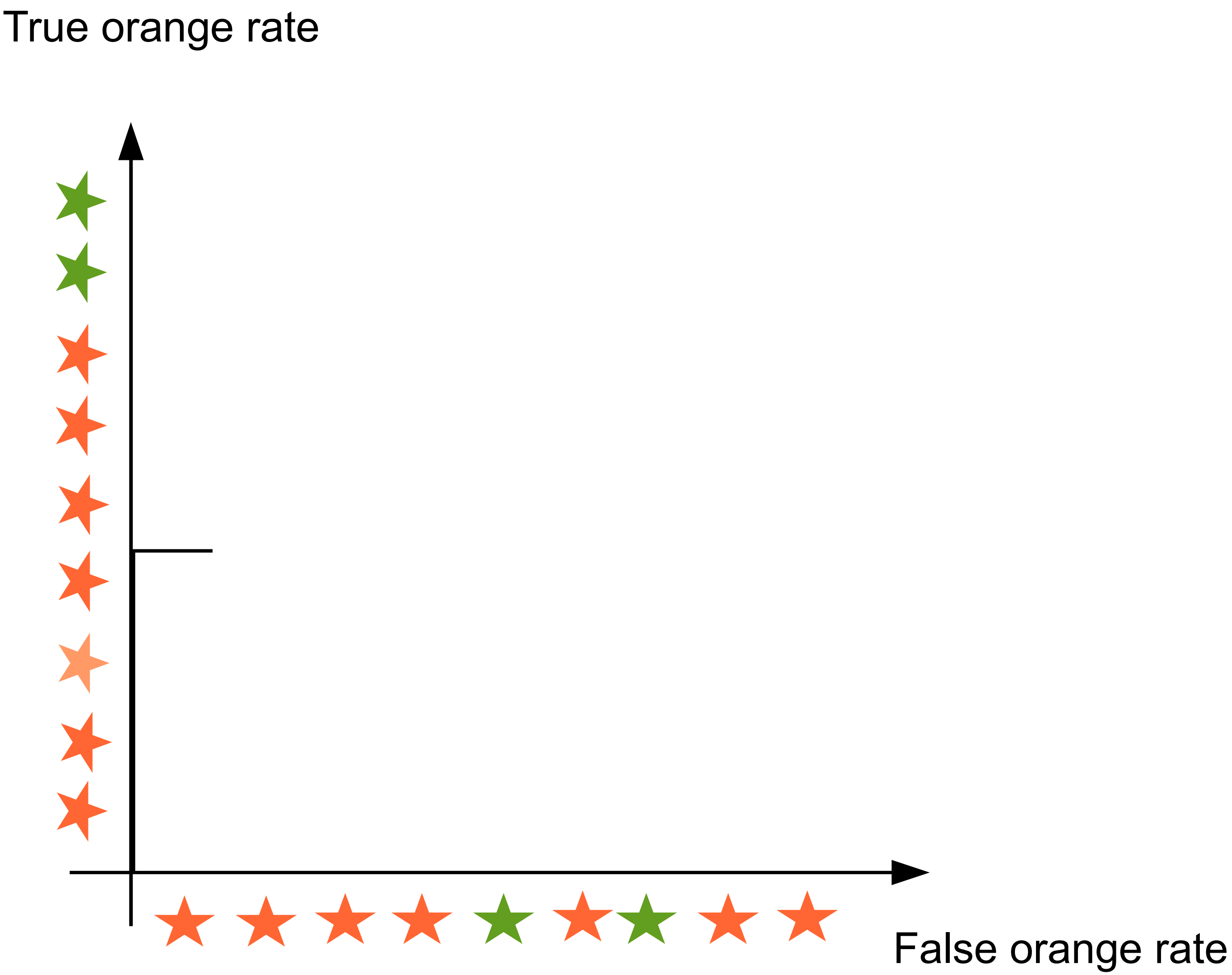
ROC Curves



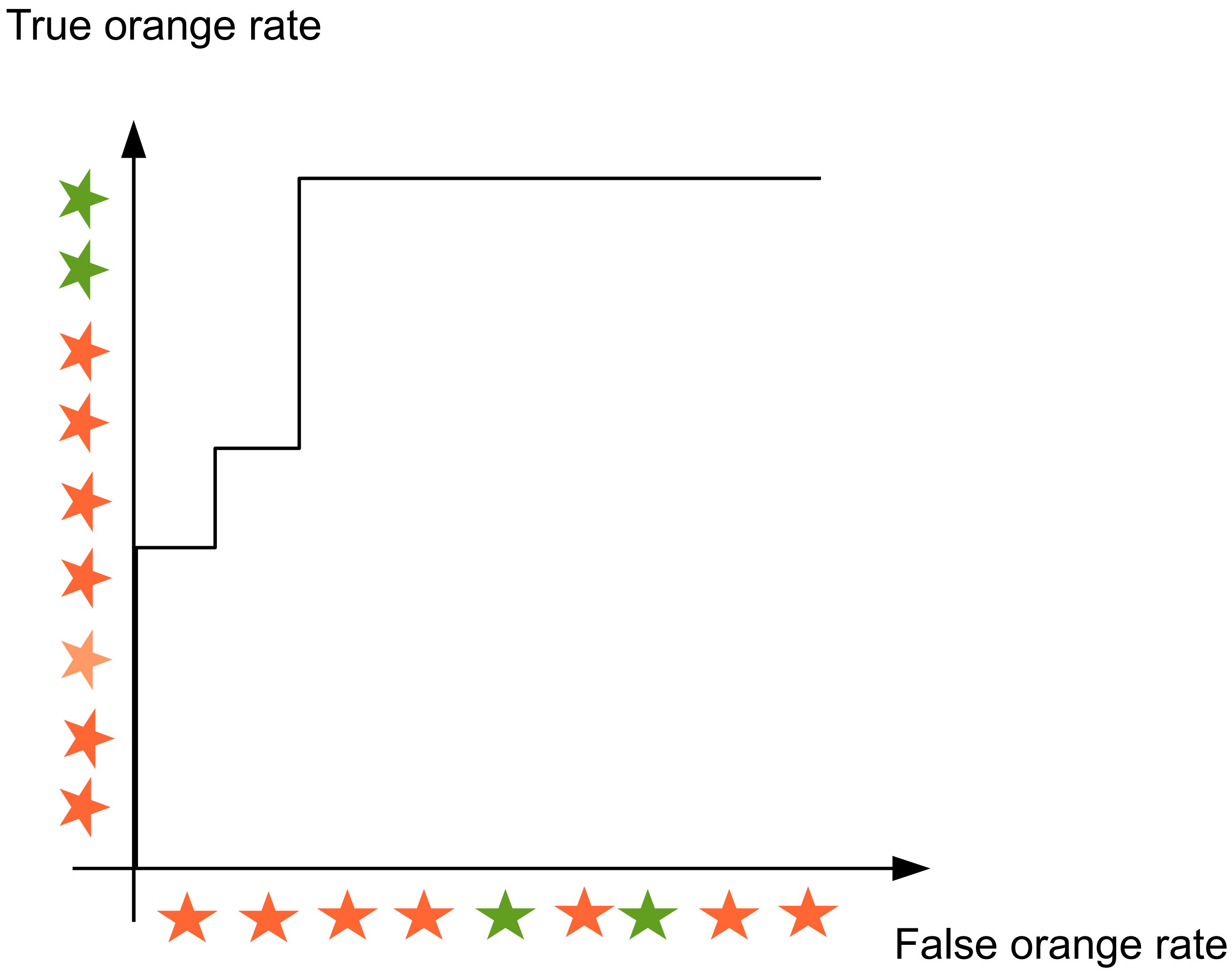
ROC Curves



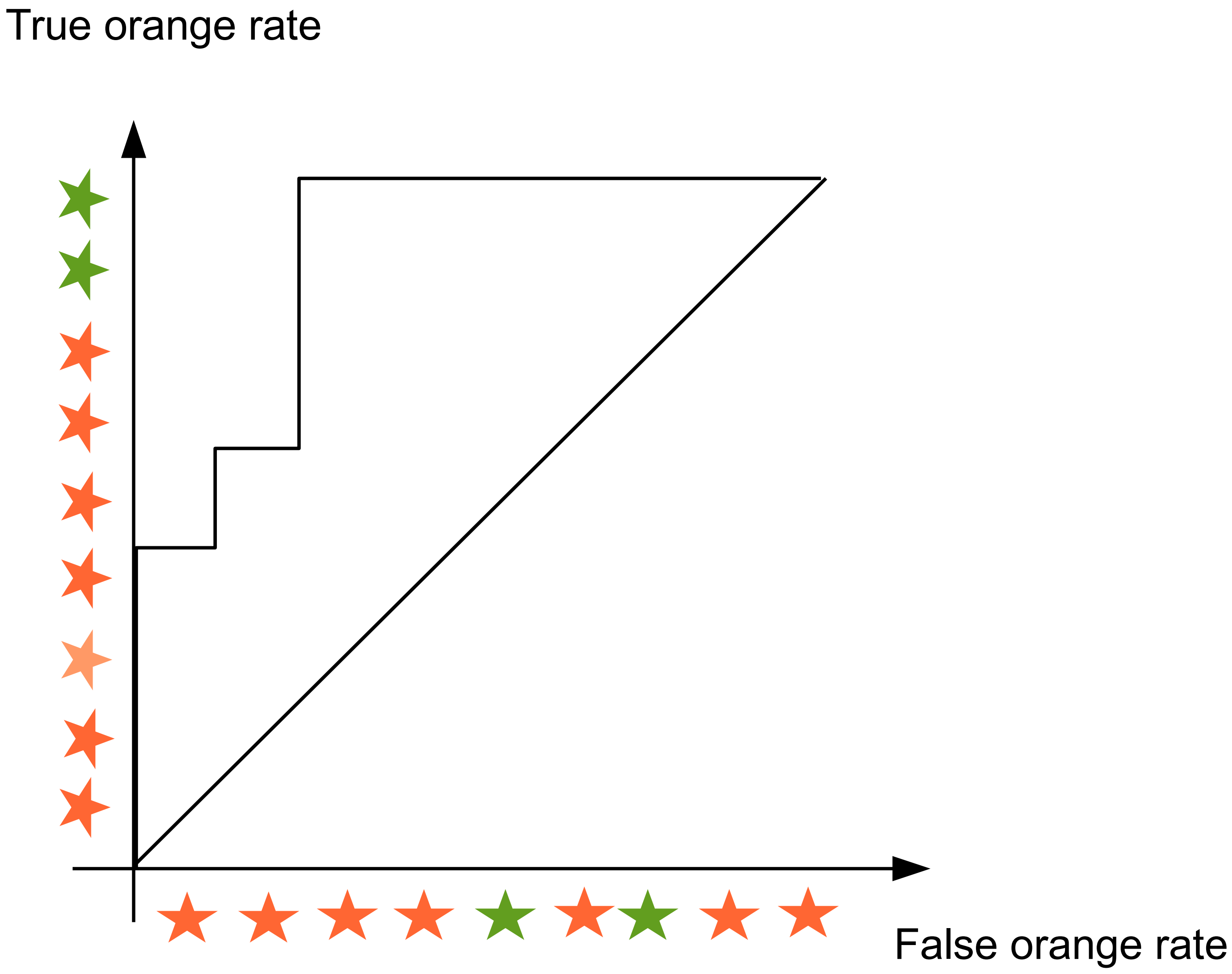
ROC Curves



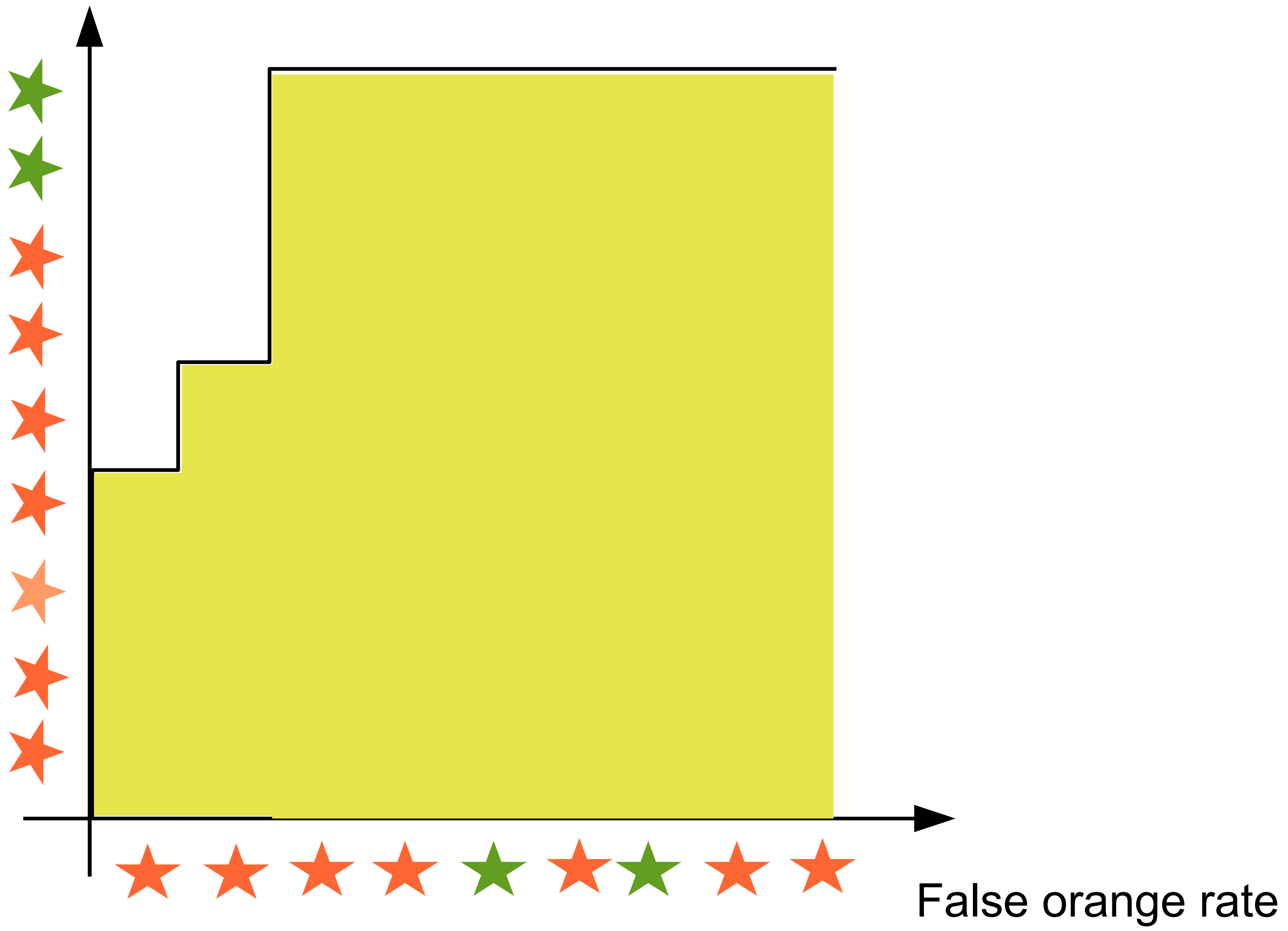
ROC Curves



ROC Curves

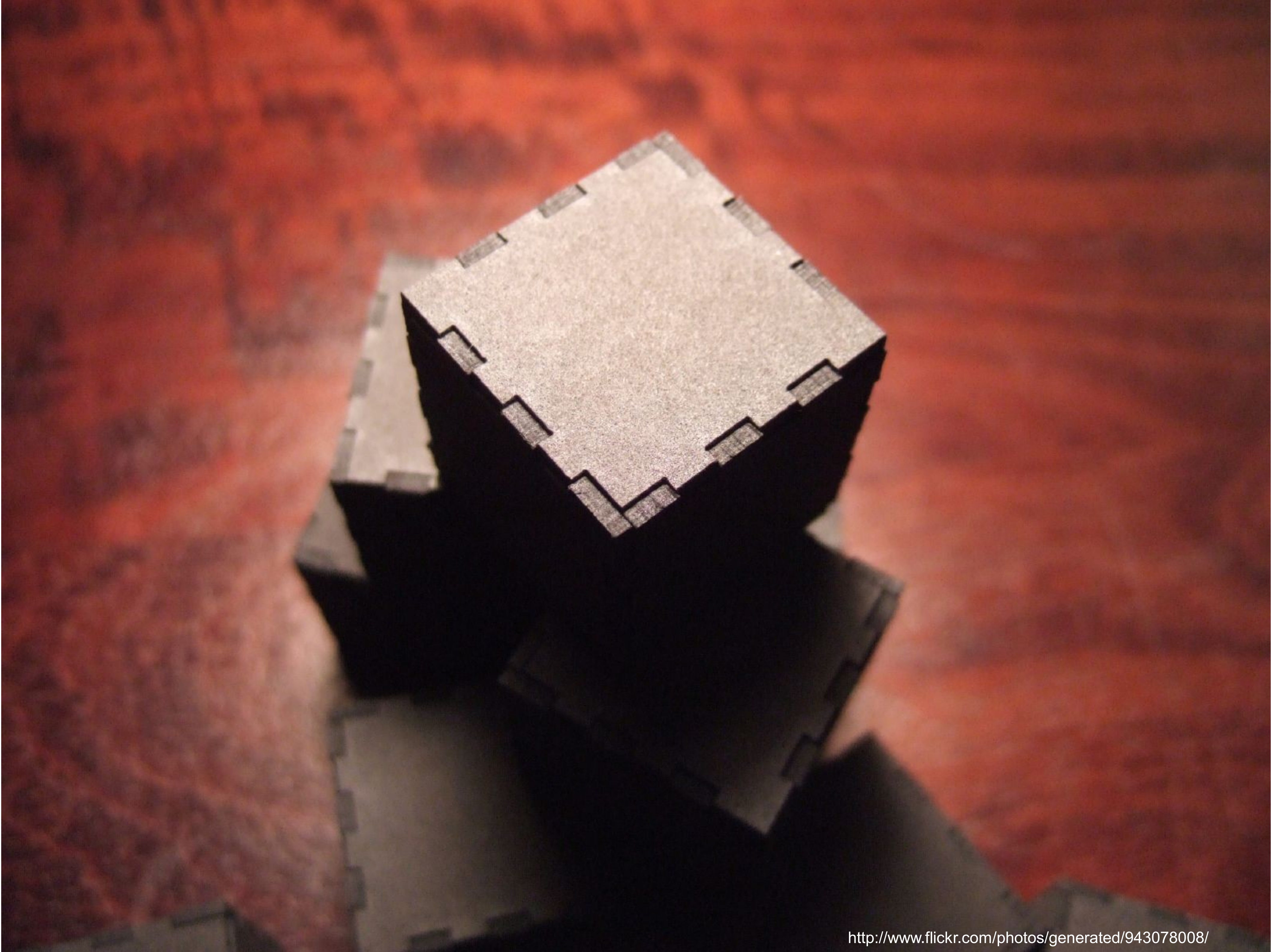


AUC – area under ROC



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

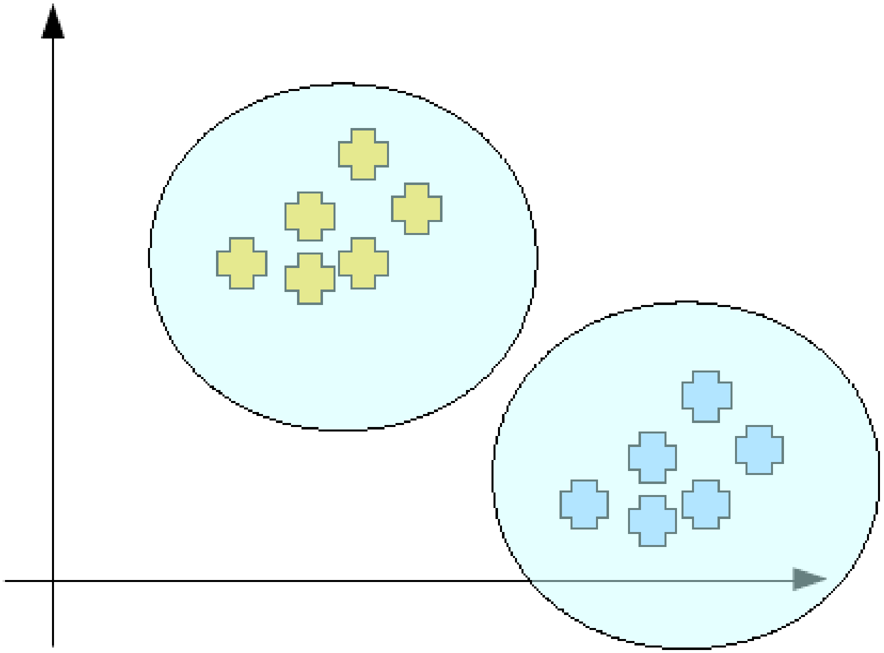
- Given a dataset of texts, identify main topics.

Algorithms: Parallel LDA

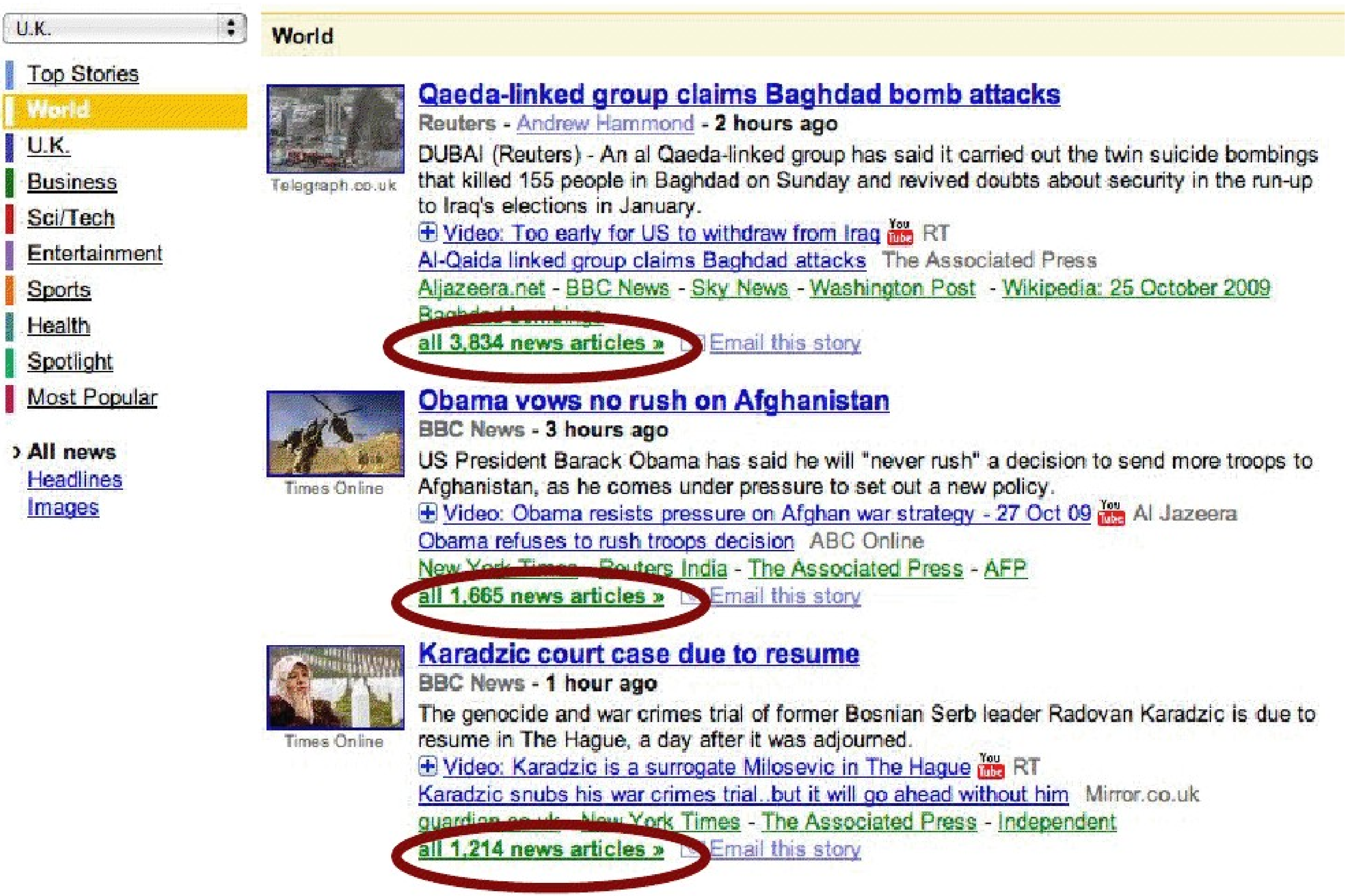
- Examples:
  - Dominant topics in set of mails.
  - Identify news message categories.

Discover groups of items

- Group items by similarity.



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



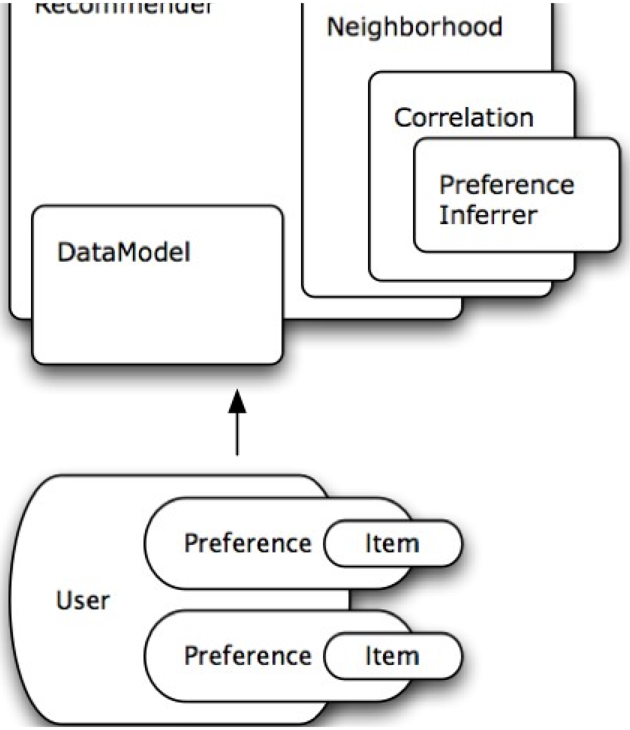
Qaeda-linked group claims Baghdad bomb attacks
Reuters - Andrew Hammond - **2 hours ago**
DUBAI (Reuters) - An al Qaeda-linked group has said it carried out the twin suicide bombings that killed 155 people in Baghdad on Sunday and revived doubts about security in the run-up to Iraq's elections in January.
⊞ 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

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

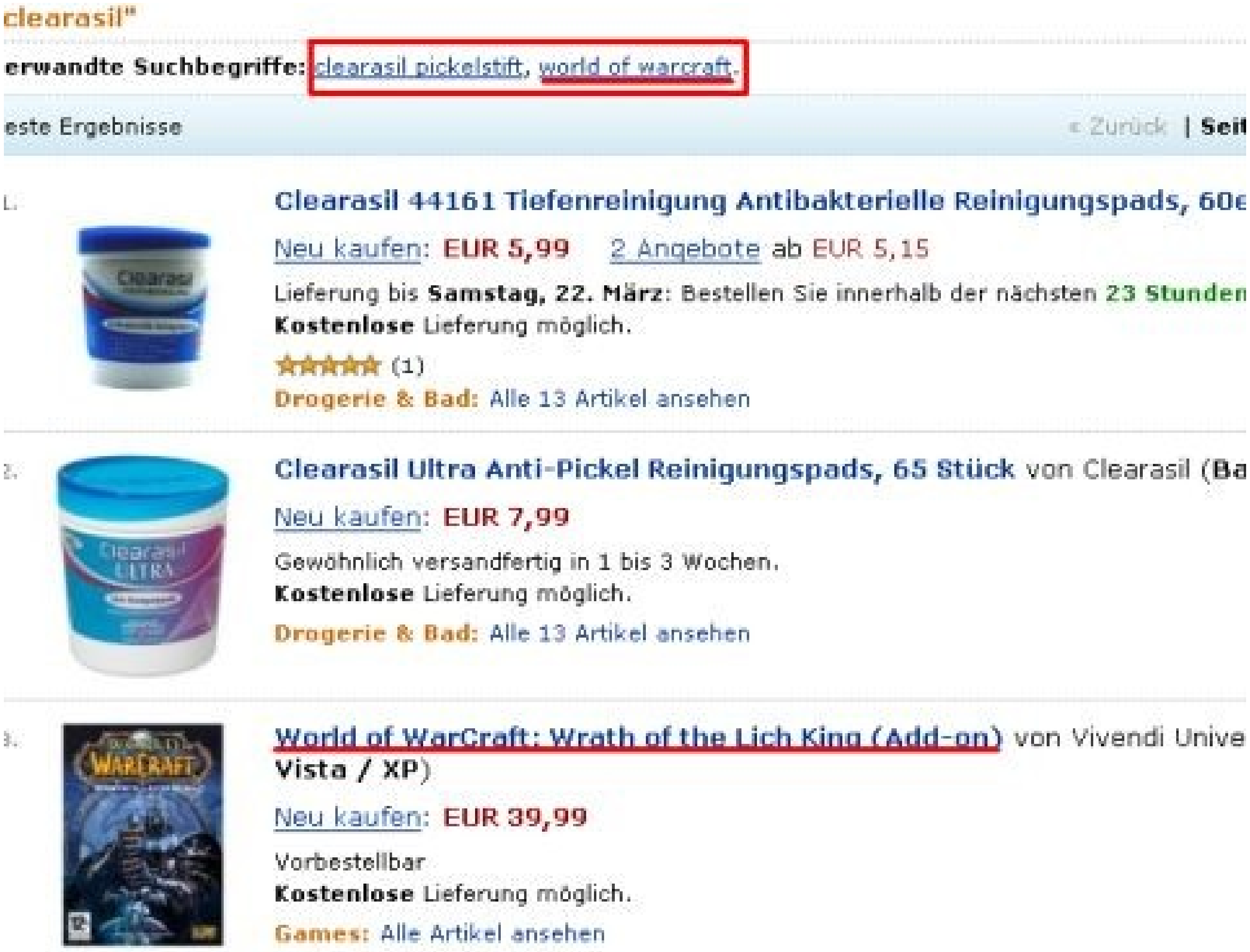
Karadzic court case due to resume
BBC News - **1 hour ago**
The genocide and war crimes trial of former Bosnian Serb leader Radovan Karadzic is due to resume in The Hague, a day after it was adjourned.
⊞ Video: Karadzic is a surrogate Milosevic in The Hague RT
Karadzic snubs his war crimes trial..but 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

Recommendation mining.

- Collaborative filtering.



Show most relevant ads



clearasil"

erwandte Suchbegriffe: clearasil pickelstift, world of warcraft.

este Ergebnisse

« Zurück | Seit

1. Clearasil 44161 Tiefenreinigung Antibakterielle Reinigungspads, 60e

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)

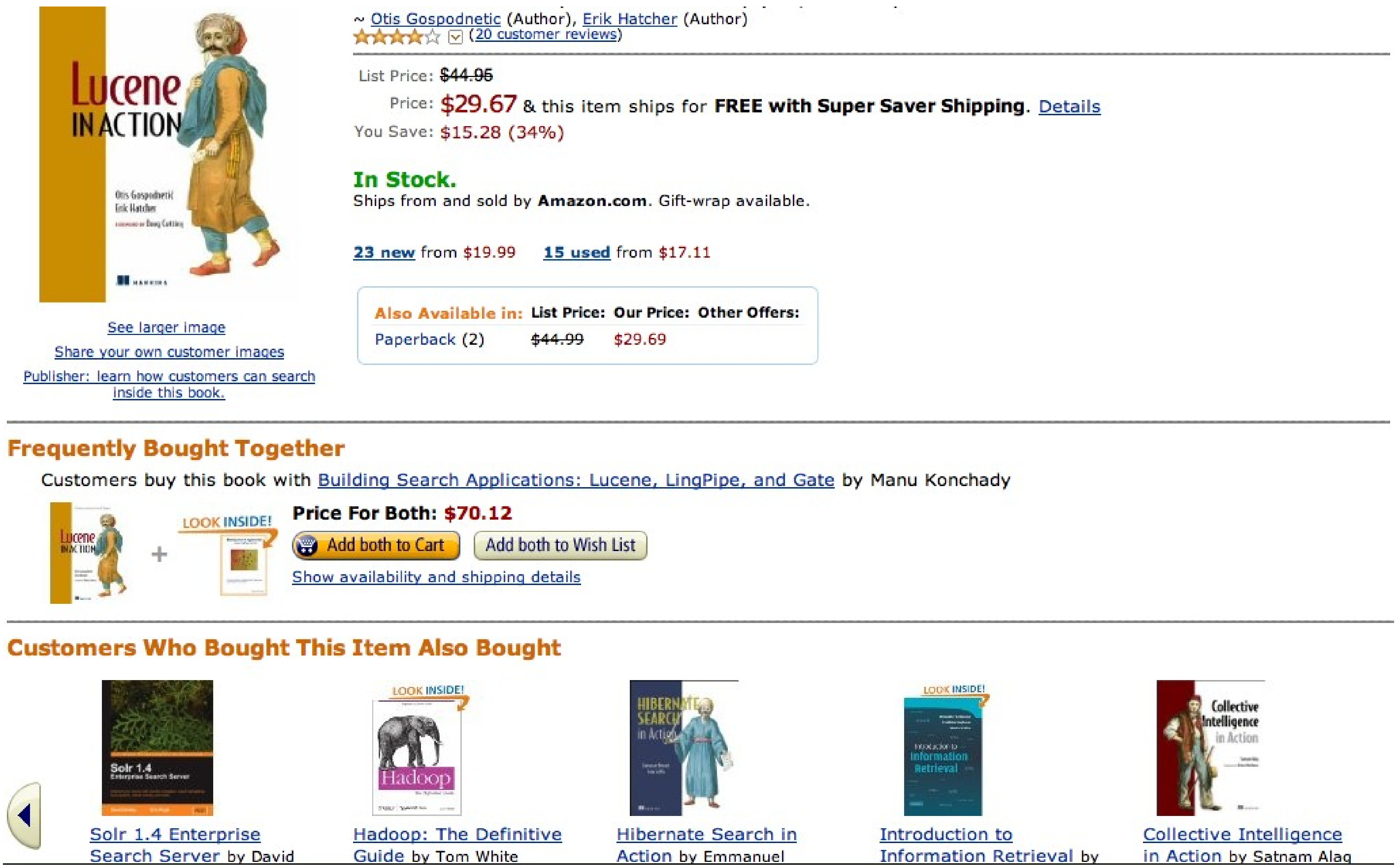
Neu kaufen: EUR 39,99

Vorbestellbar

Kostenlose Lieferung möglich.

Games: Alle Artikel ansehen

Show most relevant ads



~ 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

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

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

http://www.flickr.com/photos/25831000@N08/4156701164

http://www.flickr.com/photos/alainpicard/4175214747

http://www.flickr.com/photos/joachim_s_mueller/2417313476/

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

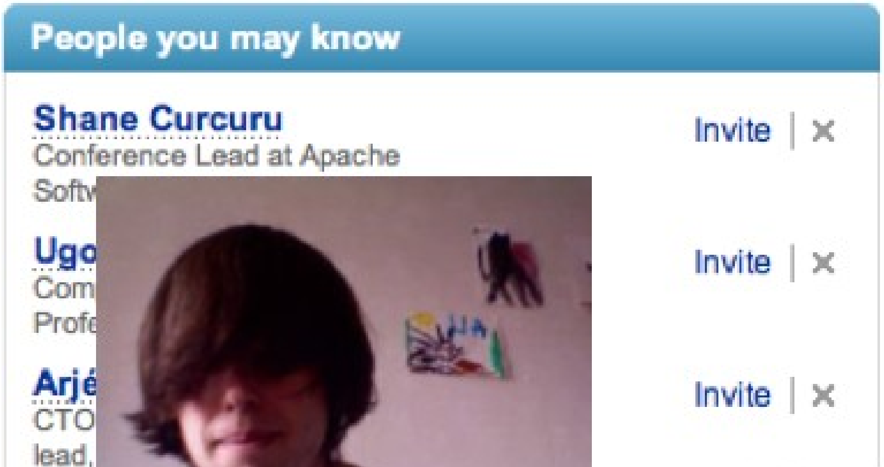
http://www.flickr.com/photos/philfotos/4510197138/

http://www.flickr.com/photos/claudio_ar/2643165035/

http://www.flickr.com/photos/claudio_ar/2643180457

Thanks to Falko Menge for the pictures of Brussels.

Recommending people



People you may know

Shane Curcuru
Conference Lead at Apache
Softw

Ugo
Com
Profe

Arjé
CTO
lead,
Mem
Foun

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.

rypto/3201254932/sizes/l/

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

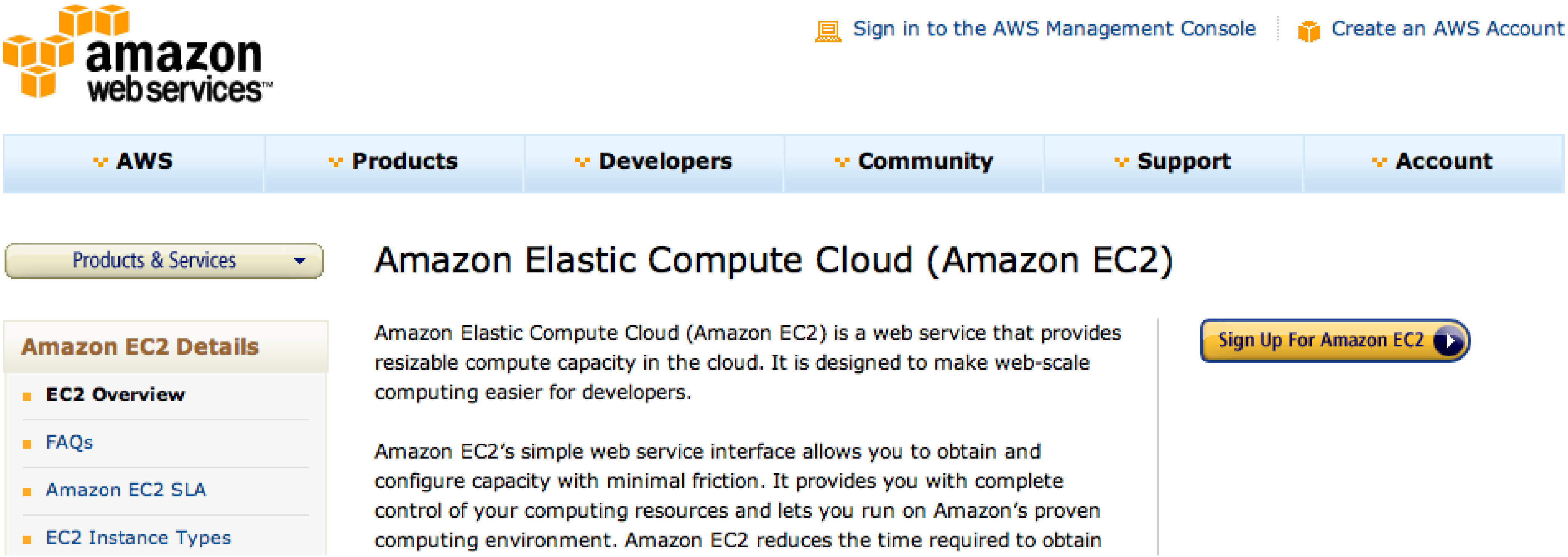
By quinnanya, http://www.flickr.com/photos/quinnanya/2806883231/

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

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

Requirements to get started

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



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.

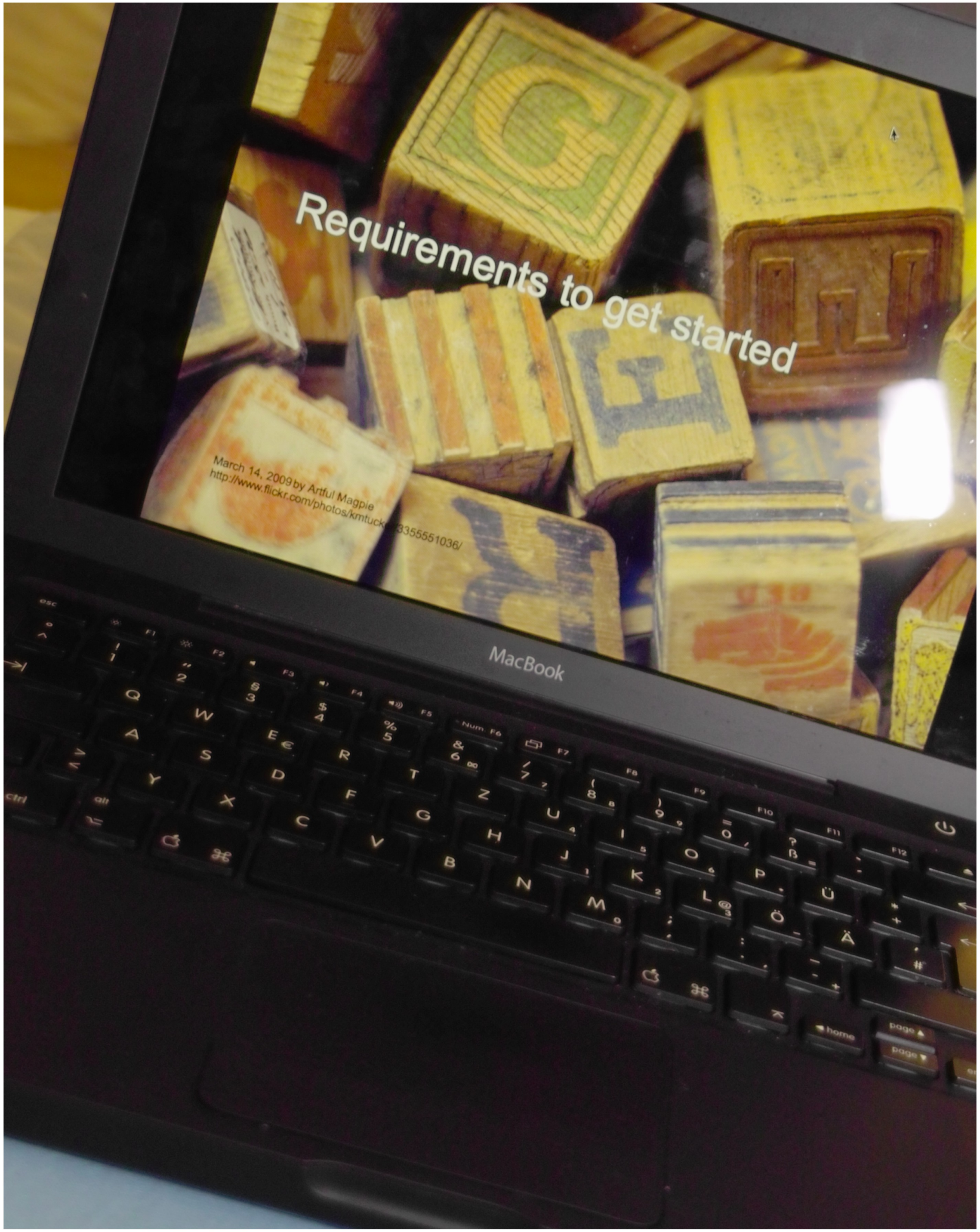
Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain

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/kmtuck 3355551036/

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.

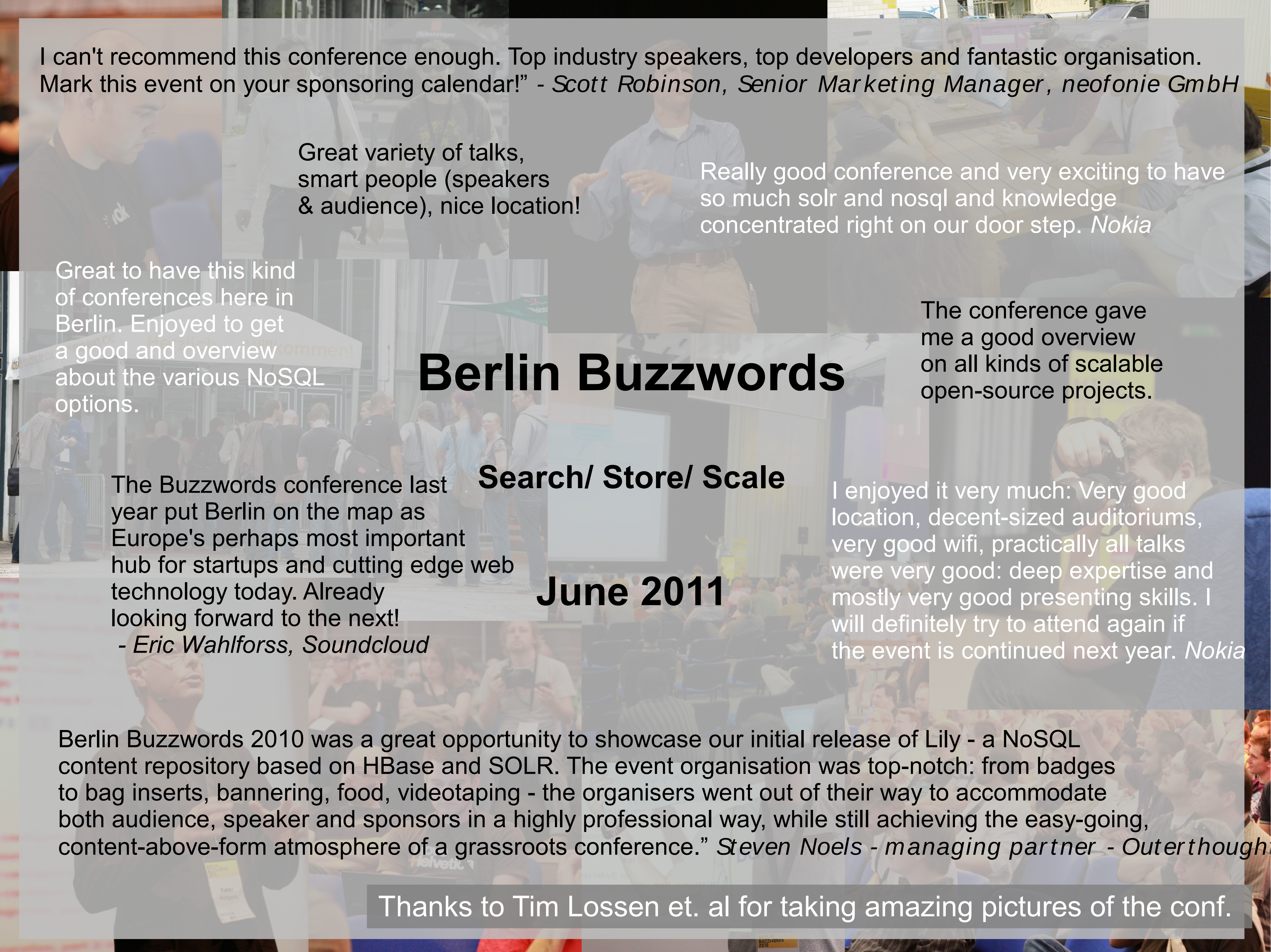
Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.

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*

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*

June 2011

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 - Outerthough*

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.