Apache Mahout

Scaling Machine Learning

Presented by:
Isabel Drost

Agenda

- Motivation.
- Machine learning?
- Introducing Mahout.
- How can you help?

Some motivation.

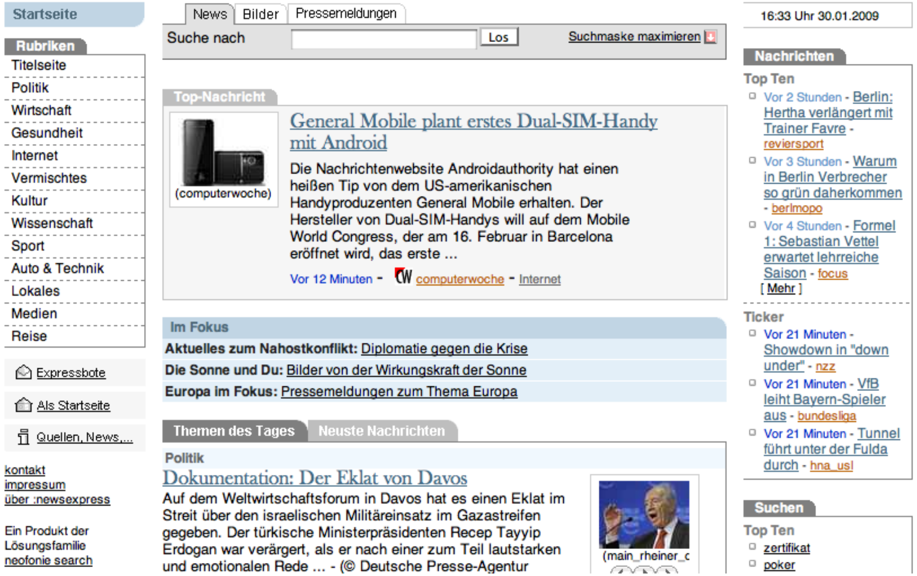
COMMUNITY NEWS
Finishing touches still to come
A glimpse of today, yesterday

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

Follow news stories

September 10, 2008 by Alex Barth
http://www.flickr.com/photos/a-barth/2846621384

Search through papers.



Startseite

Rubriken

- Titelseite
- Politik
- Wirtschaft
- Gesundheit
- Internet
- Vermischtes
- Kultur
- Wissenschaft
- Sport
- Auto & Technik
- Lokales
- Medien
- Reise

Expressbote

Als Startseite

Quellen, News,...

kontakt
impressum
über :newsexpress

Ein Produkt der
Lösungsfamilie
neofonie search

News | Bilder | Pressemeldungen

Suche nach [ ] Los — Suchmaske maximieren

Top-Nachricht

(computerwoche)

General Mobile plant erstes Dual-SIM-Handy mit Android

Die Nachrichtenwebsite Androidauthority hat einen heißen Tip von dem US-amerikanischen Handyproduzenten General Mobile erhalten. Der Hersteller von Dual-SIM-Handys will auf dem Mobile World Congress, der am 16. Februar in Barcelona eröffnet wird, das erste ...

Vor 12 Minuten - CW computerwoche - Internet

Im Fokus

Aktuelles zum Nahostkonflikt: Diplomatie gegen die Krise

Die Sonne und Du: Bilder von der Wirkungskraft der Sonne

Europa im Fokus: Pressemeldungen zum Thema Europa

Themen des Tages | Neuste Nachrichten

Politik

Dokumentation: Der Eklat von Davos

Auf dem Weltwirtschaftsforum in Davos hat es einen Eklat im Streit über den israelischen Militäreinsatz im Gazastreifen gegeben. Der türkische Ministerpräsidenten Recep Tayyip Erdogan war verärgert, als er nach einer zum Teil lautstarken und emotionalen Rede ... - (© Deutsche Presse-Agentur

(main_rheiner_c

16:33 Uhr 30.01.2009

Nachrichten

Top Ten

- Vor 2 Stunden - Berlin: Hertha verlängert mit Trainer Favre - reviersport
- Vor 3 Stunden - Warum in Berlin Verbrecher so grün daherkommen - berlmopo
- Vor 4 Stunden - Formel 1: Sebastian Vettel erwartet lehrreiche Saison - focus

[Mehr]

Ticker

- Vor 21 Minuten - Showdown in "down under" - nzz
- Vor 21 Minuten - VfB leiht Bayern-Spieler aus - bundesliga
- Vor 21 Minuten - Tunnel führt unter der Fulda durch - hna_usl

Suchen

Top Ten

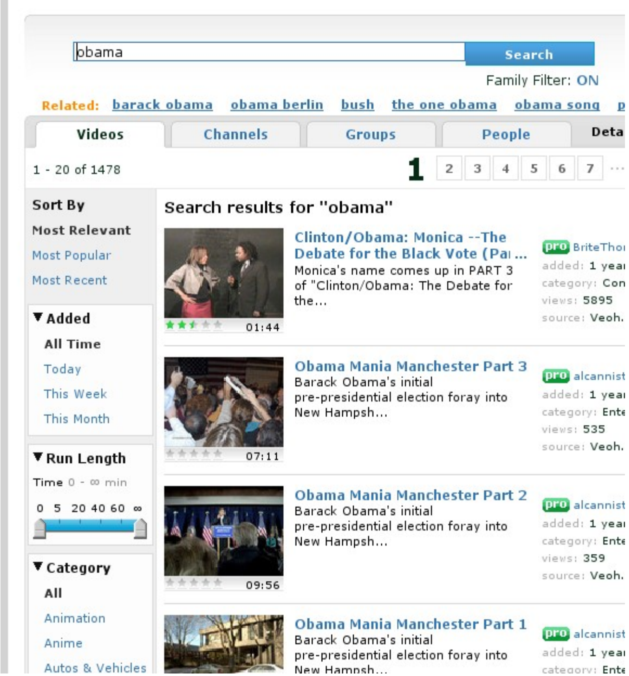
- zertifikat
- poker

Automatic topic tracker.

March 7, 2008 by extranoise
http://www.flickr.com/photos/extranoise/2317950586/

Movie recommendation

March 22, 2008 by Crystian Cruz
http://www.flickr.com/photos/crystiancruz/2353895708



obama Search

Family Filter: ON

Related: barack obama obama berlin bush the one obama obama song p

Videos | Channels | Groups | People | Deta

1 - 20 of 1478 — 1 2 3 4 5 6 7 ...

Sort By
Most Relevant
Most Popular
Most Recent

▼Added
All Time
Today
This Week
This Month

▼Run Length
Time 0 - ∞ min
0 5 20 40 60 ∞

▼Category
All
Animation
Anime
Autos & Vehicles

Search results for "obama"

Clinton/Obama: Monica --The Debate for the Black Vote (Pa ... — 01:44
Monica's name comes up in PART 3 of "Clinton/Obama: The Debate for the...
pro BriteTho
added: 1 yea
category: Con
views: 5895
source: Veoh.

Obama Mania Manchester Part 3 — 07:11
Barack Obama's initial pre-presidential election foray into New Hampsh...
pro alcannist
added: 1 yea
category: Ente
views: 535
source: Veoh.

Obama Mania Manchester Part 2 — 09:56
Barack Obama's initial pre-presidential election foray into New Hampsh...
pro alcannist
added: 1 yea
category: Ente
views: 359
source: Veoh.

Obama Mania Manchester Part 1
Barack Obama's initial pre-presidential election foray into New Hampsh...
pro alcannist
added: 1 yea
category: Ente

IMDB + movie reviews.

Aggregate reviews from IMDB, twitter, ...

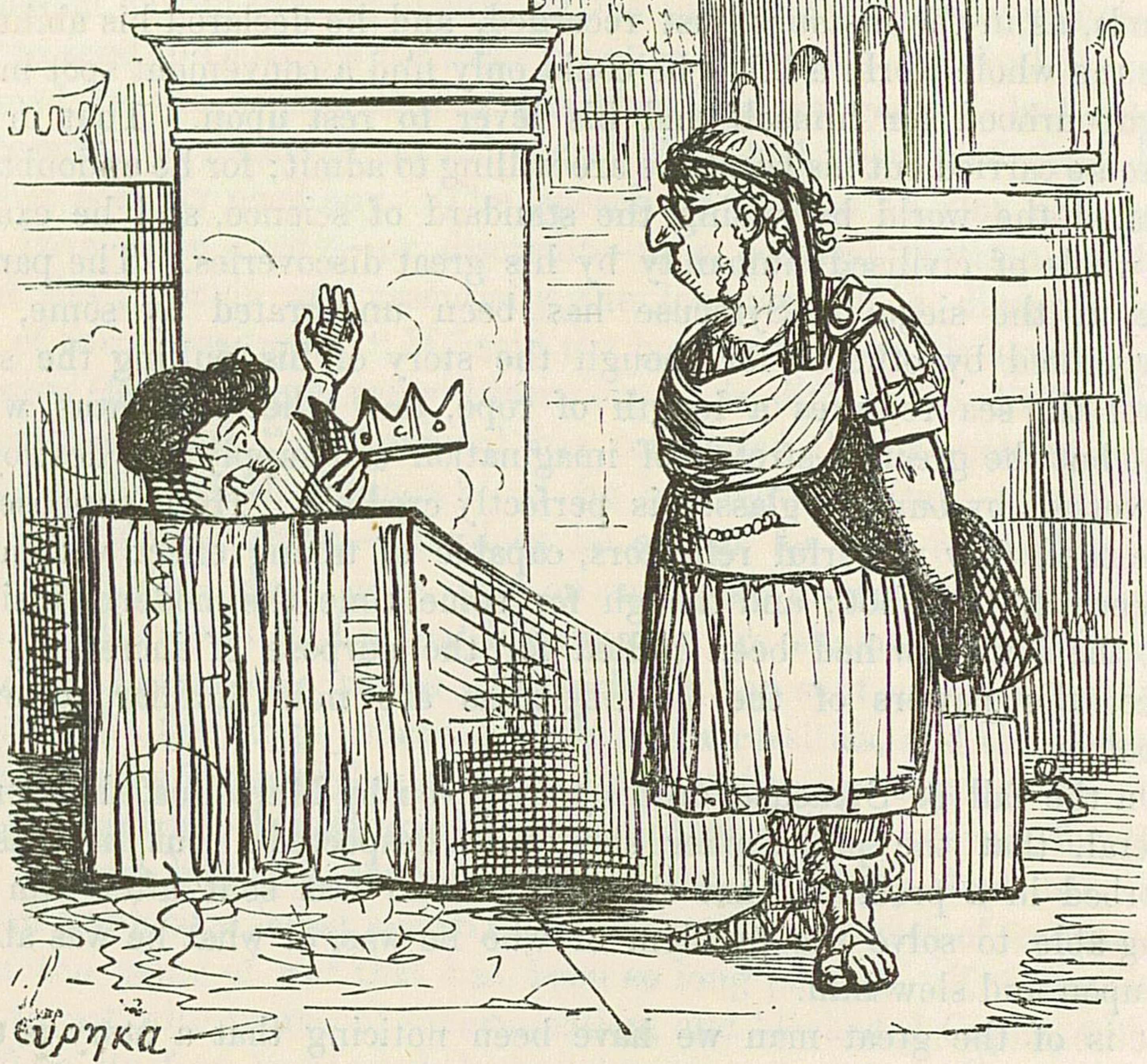
- Lots and lots of data.

- Structured and unstructured.

Mission

Provide scalable data mining algorithms.

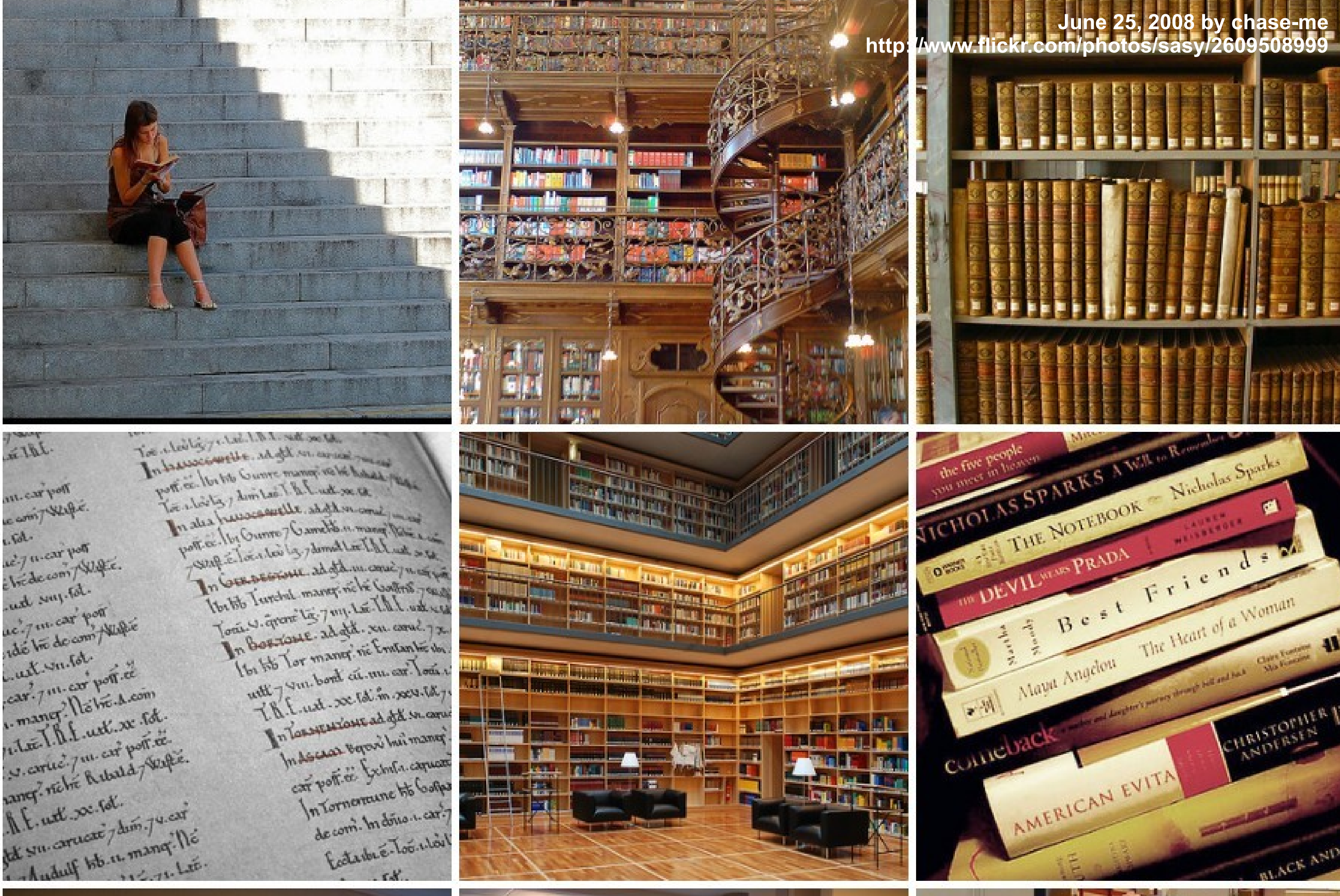
Machine Learning?


εύρηκα

Archimedes generates model:

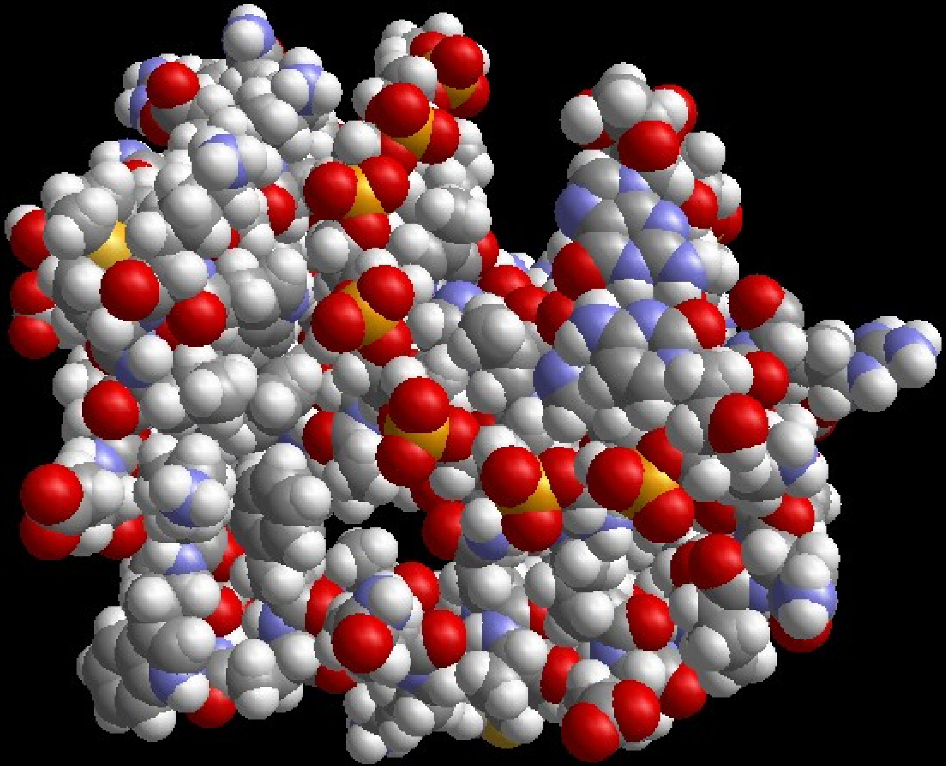
$$\frac{Density\ of\ Object}{Density\ of\ Fluid} = .$$

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

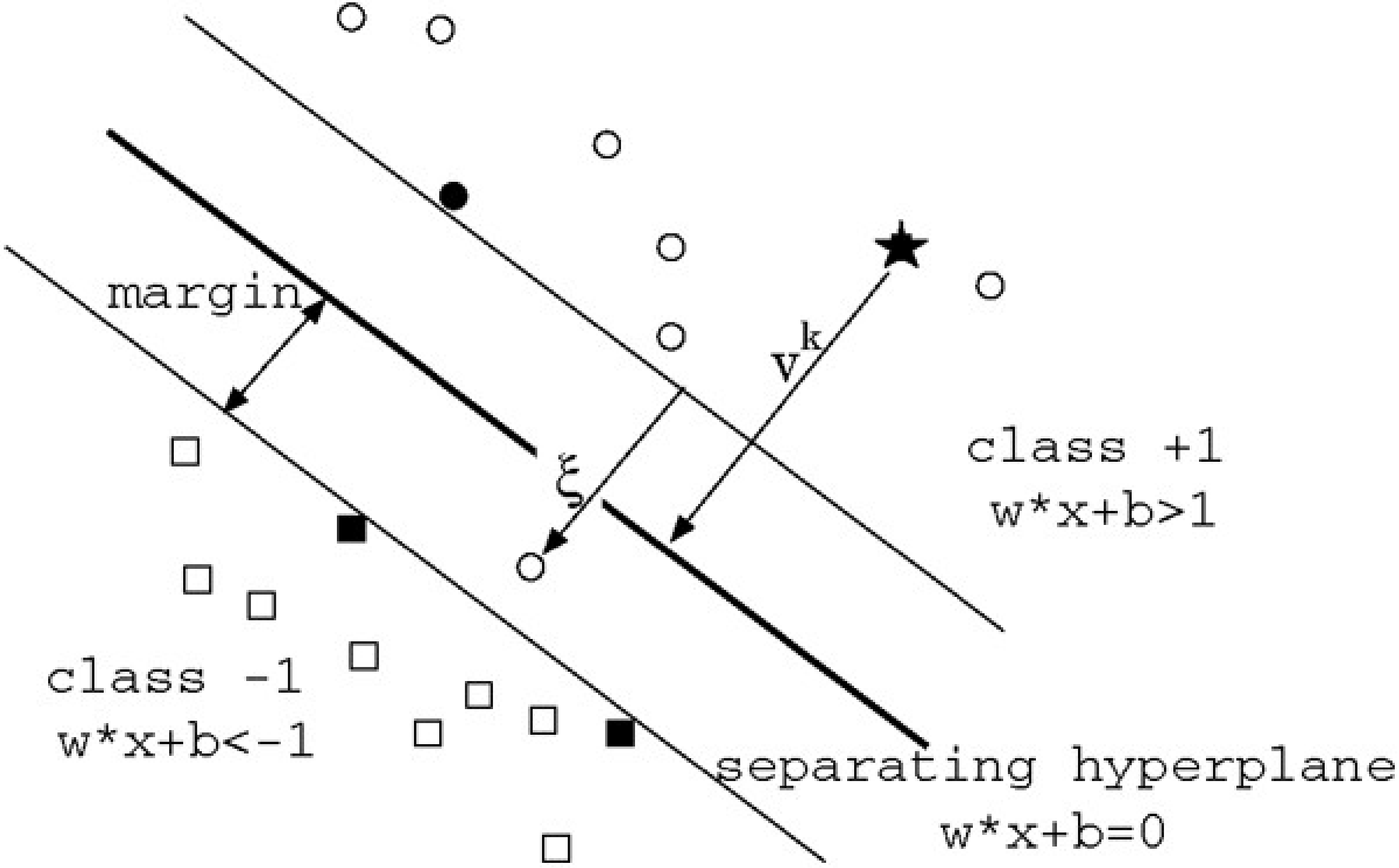


June 25, 2008 by chase-me
http://www.flickr.com/photos/sasy/2609508999

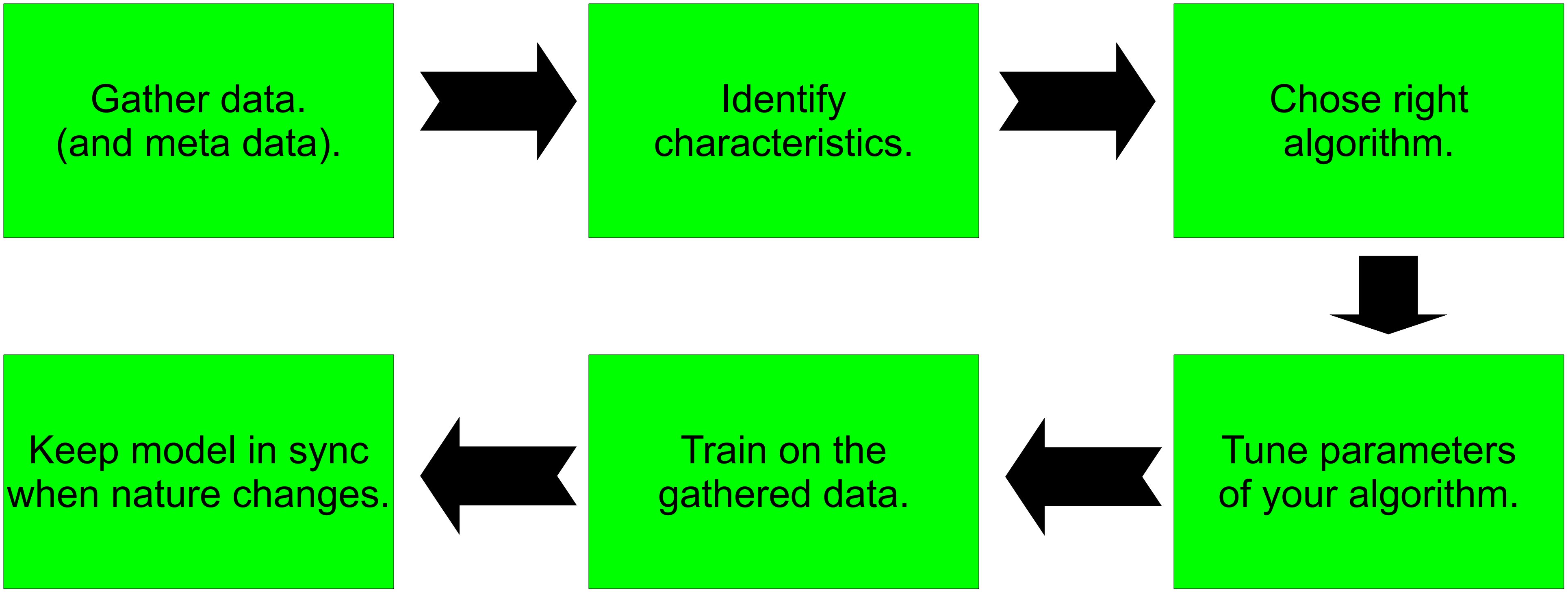
March 28, 2007 by dullhunk
http://www.flickr.com/photos/dullhunk/437551254



Machine learning generates model



Machine learning pipeline



Gather data.
(and meta data).

Identify
characteristics.

Chose right
algorithm.

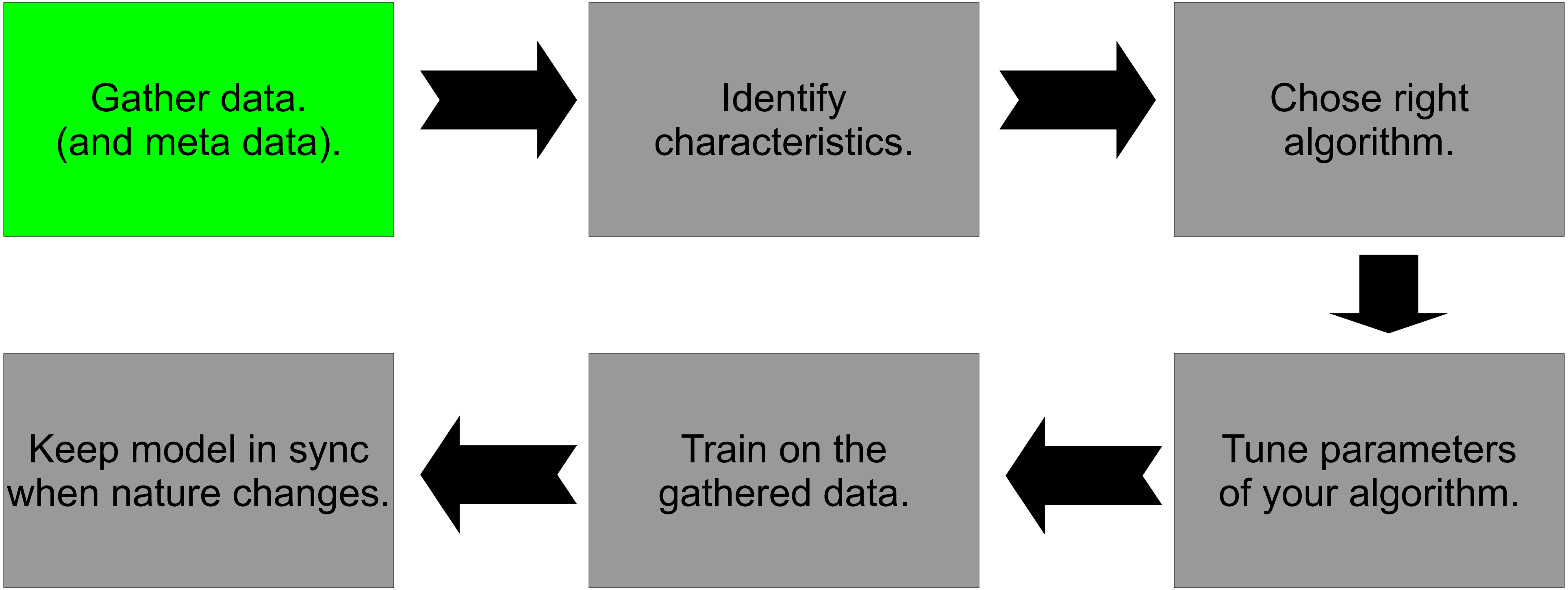
Tune parameters
of your algorithm.

Train on the
gathered data.

Keep model in sync
when nature changes.

January 8, 2008 by Pink Sherbet Photography
http://www.flickr.com/photos/pinksherbet/2177961471/

Machine learning pipeline



Gather data.
(and meta data).

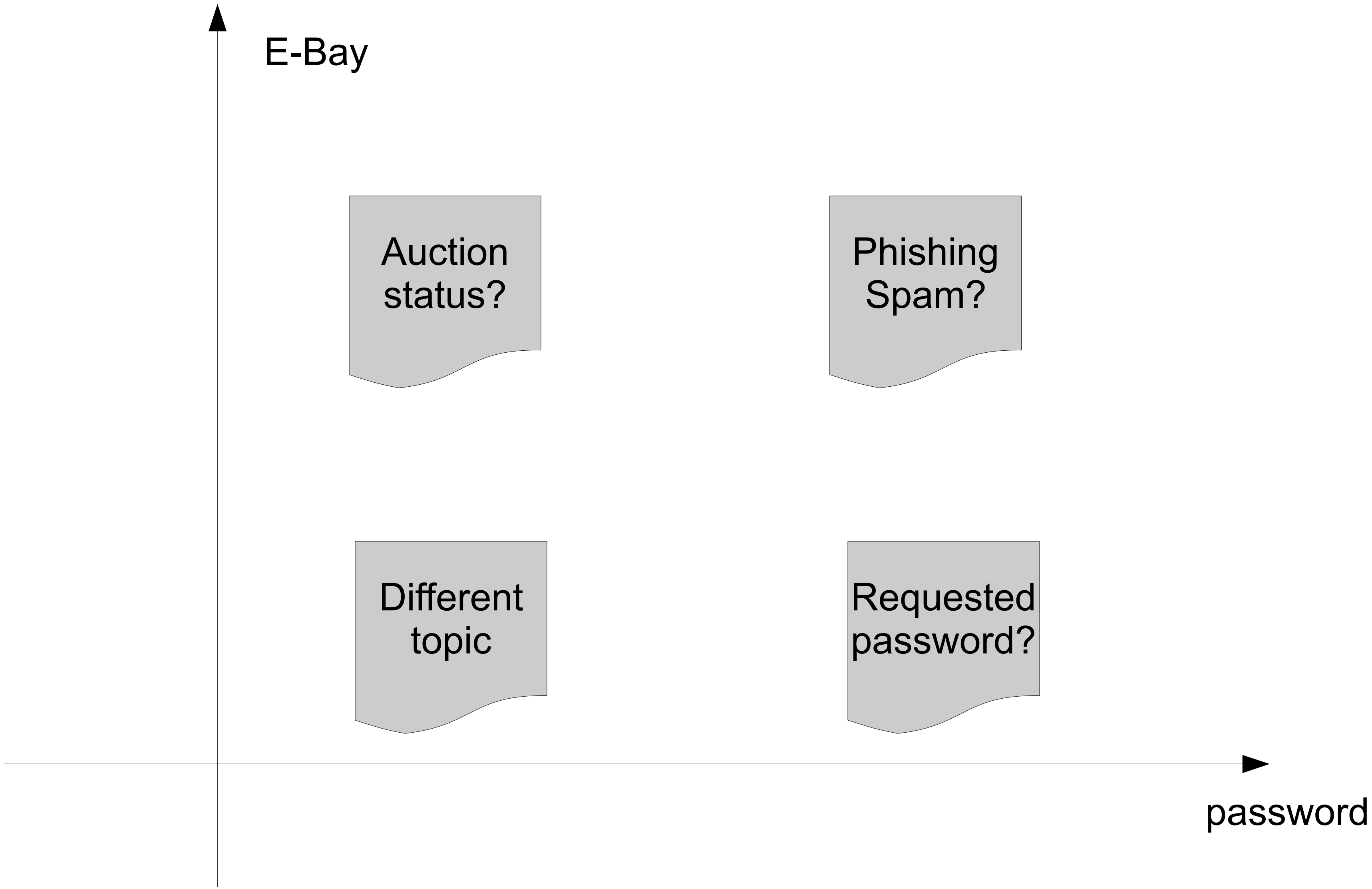
Identify
characteristics.

Chose right
algorithm.

Tune parameters
of your algorithm.

Train on the
gathered data.

Keep model in sync
when nature changes.



E-Bay

Auction status?

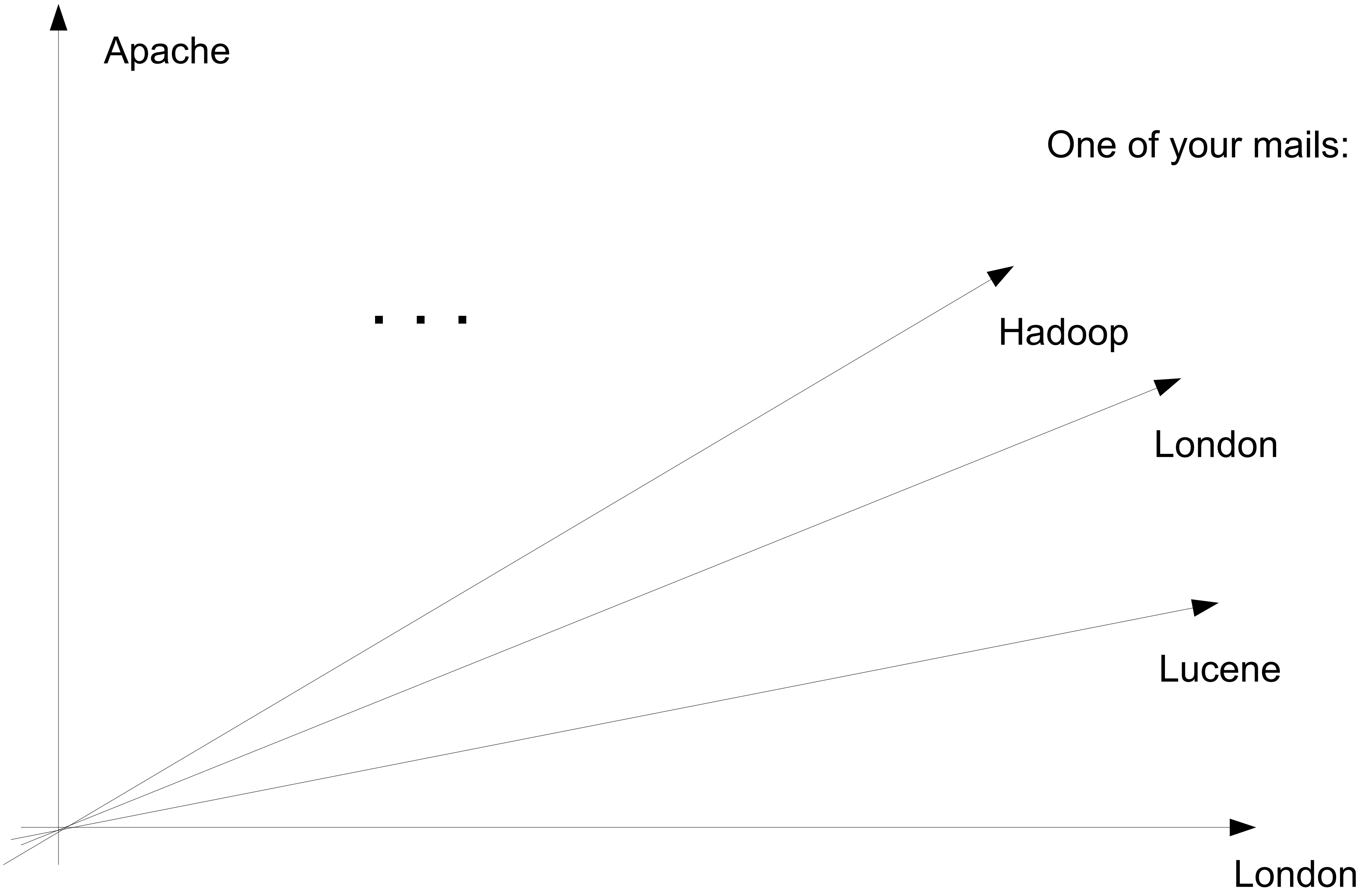
Phishing Spam?

Different topic

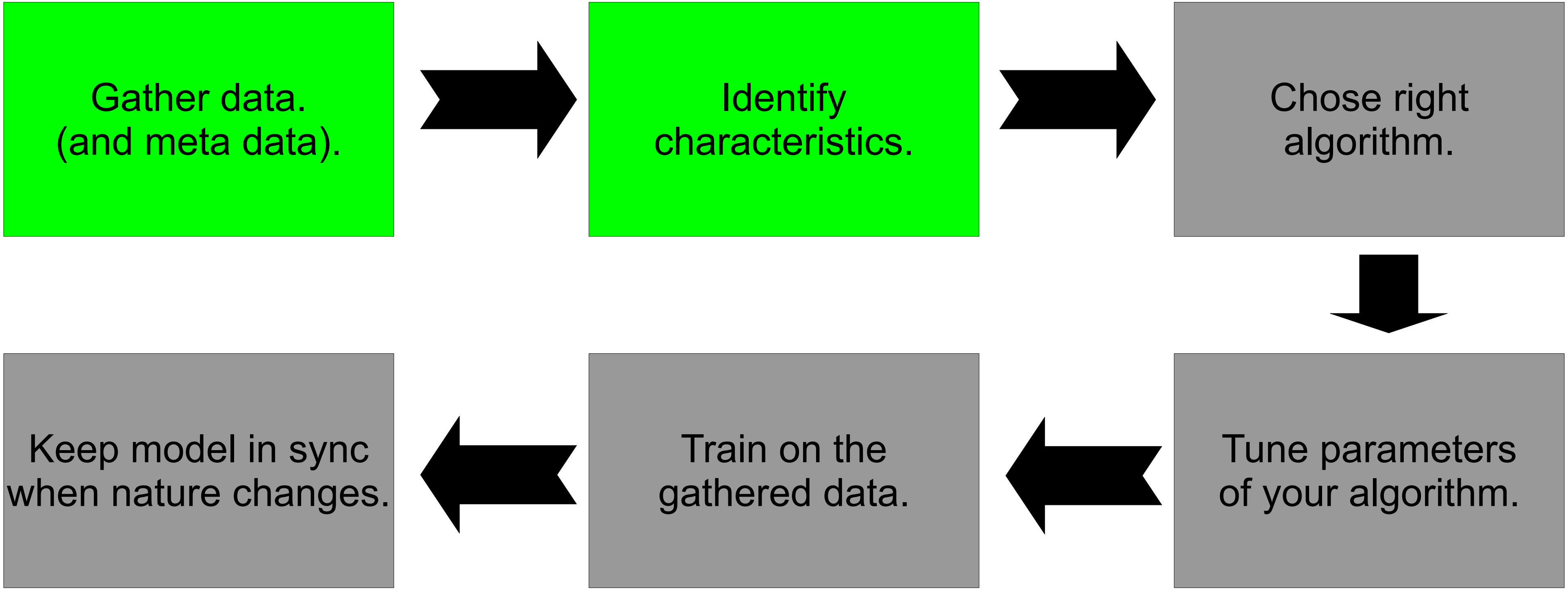
Requested password?

password

One of your mails: $\begin{pmatrix}0\\1\\10\\0\\0\\0\\1\\1\end{pmatrix}$



Machine learning pipeline



Gather data.
(and meta data).

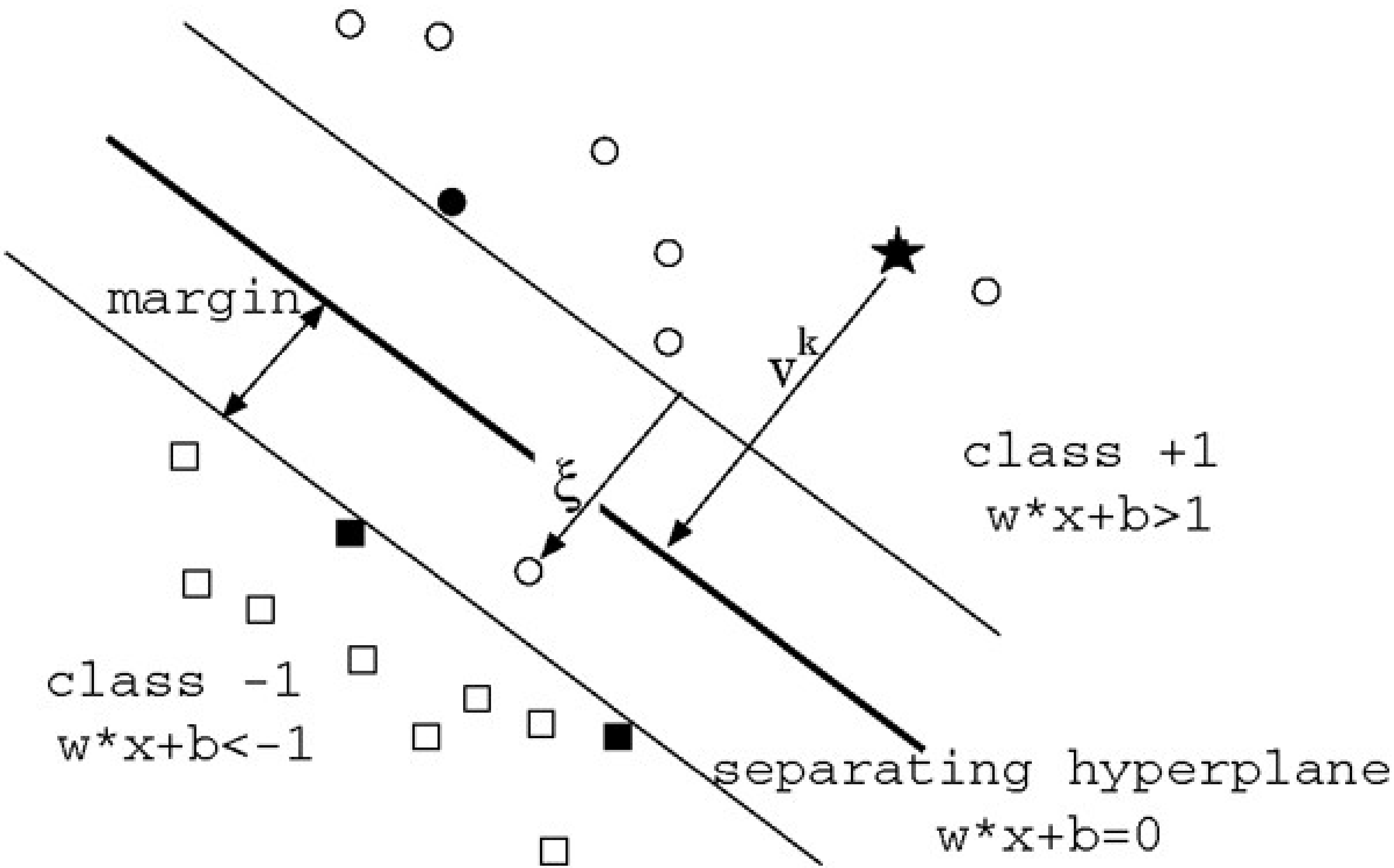
Identify
characteristics.

Chose right
algorithm.

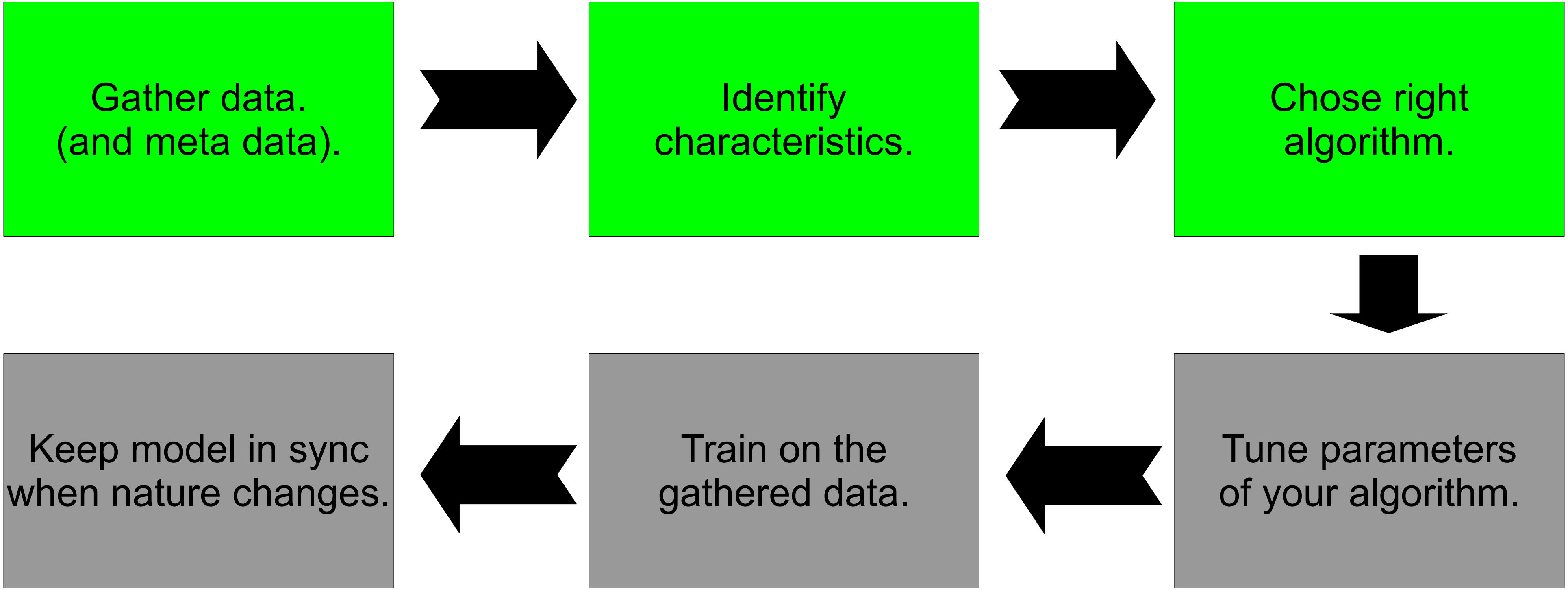
Tune parameters
of your algorithm.

Train on the
gathered data.

Keep model in sync
when nature changes.


margin
v^k
ξ
class +1
w*x+b>1
class -1
w*x+b<-1
separating hyperplane
w*x+b=0

Machine learning pipeline



Gather data. (and meta data).

Identify characteristics.

Chose right algorithm.

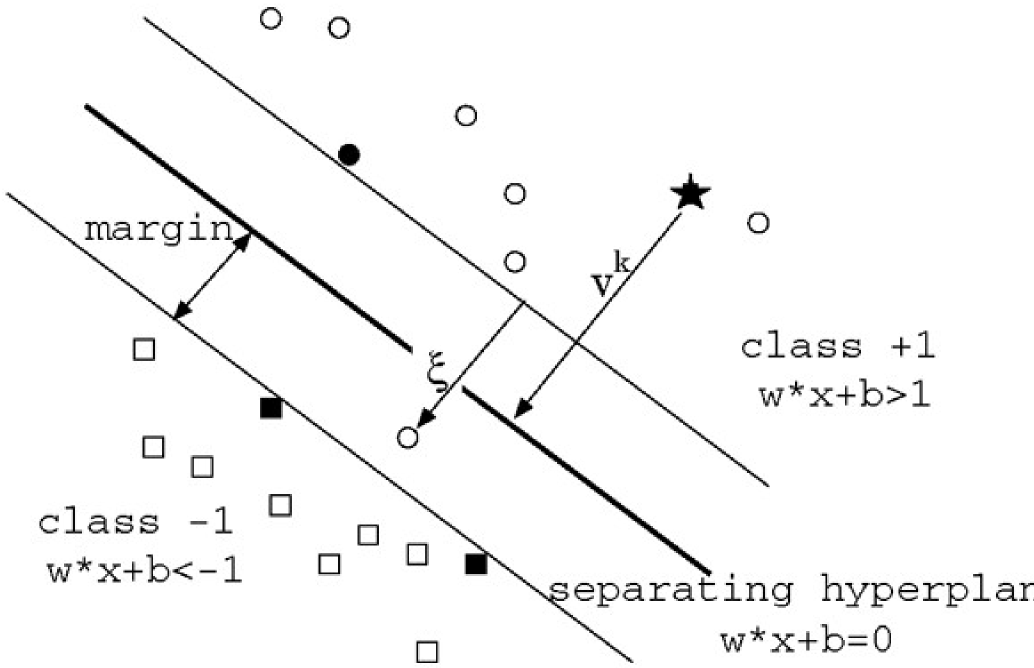
Tune parameters of your algorithm.

Train on the gathered data.

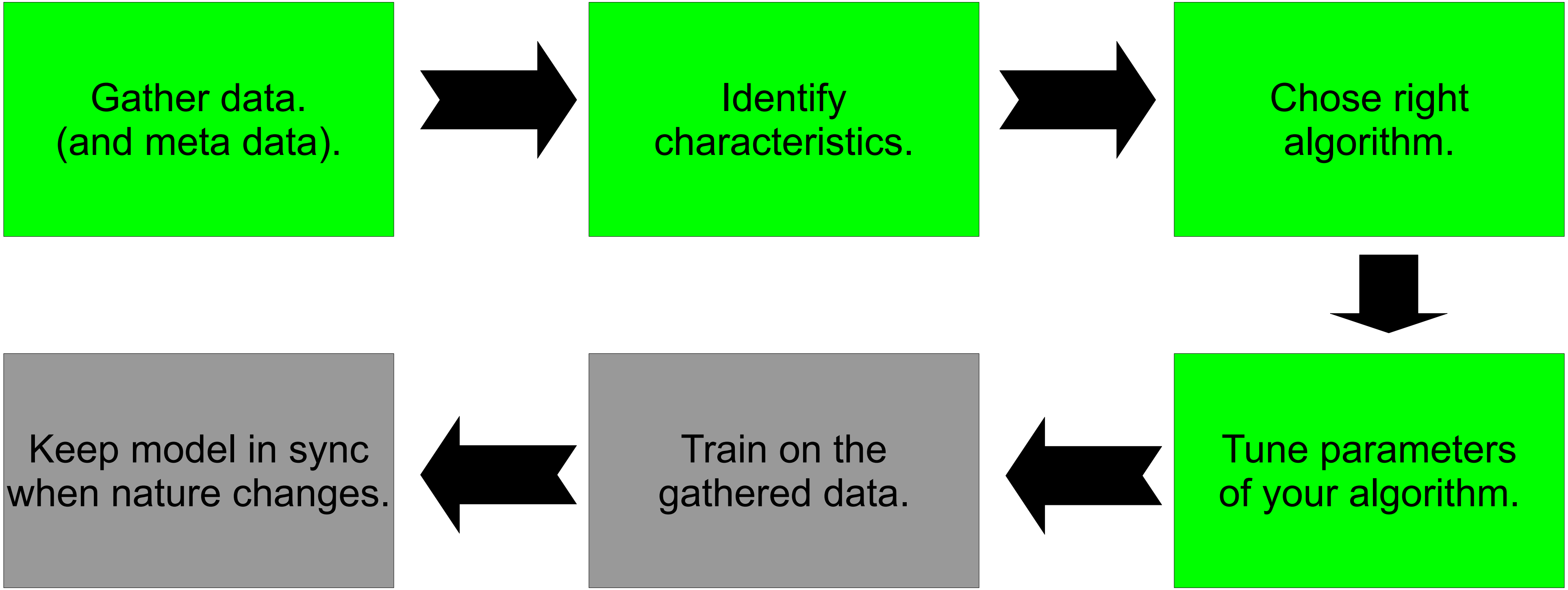
Keep model in sync when nature changes.

Parameter tuning

- Penalty for mistakes.
- Kernel type for data transformation.
- Tune kernel parameters.



Machine learning pipeline



Gather data.
(and meta data).

Identify
characteristics.

Chose right
algorithm.

Tune parameters
of your algorithm.

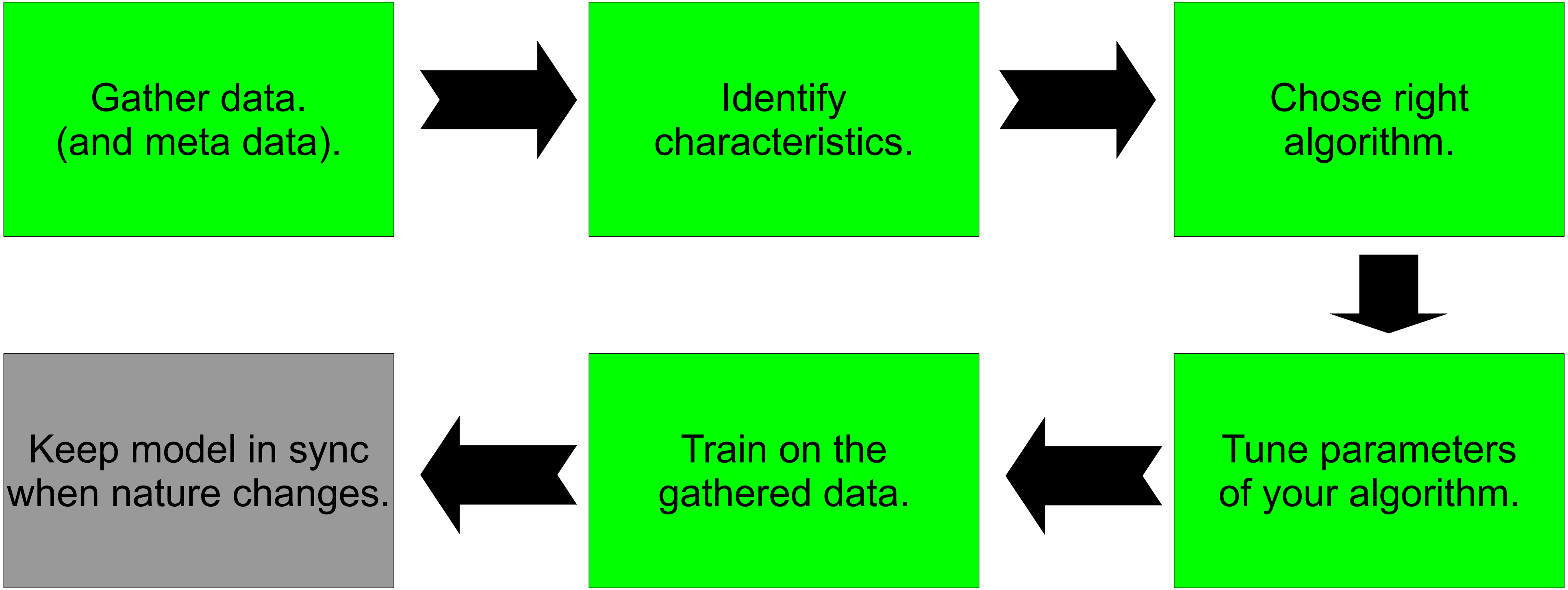
Train on the
gathered data.

Keep model in sync
when nature changes.

Training

- Build model from data.

Machine learning pipeline



Gather data.
(and meta data).

Identify
characteristics.

Chose right
algorithm.

Tune parameters
of your algorithm.

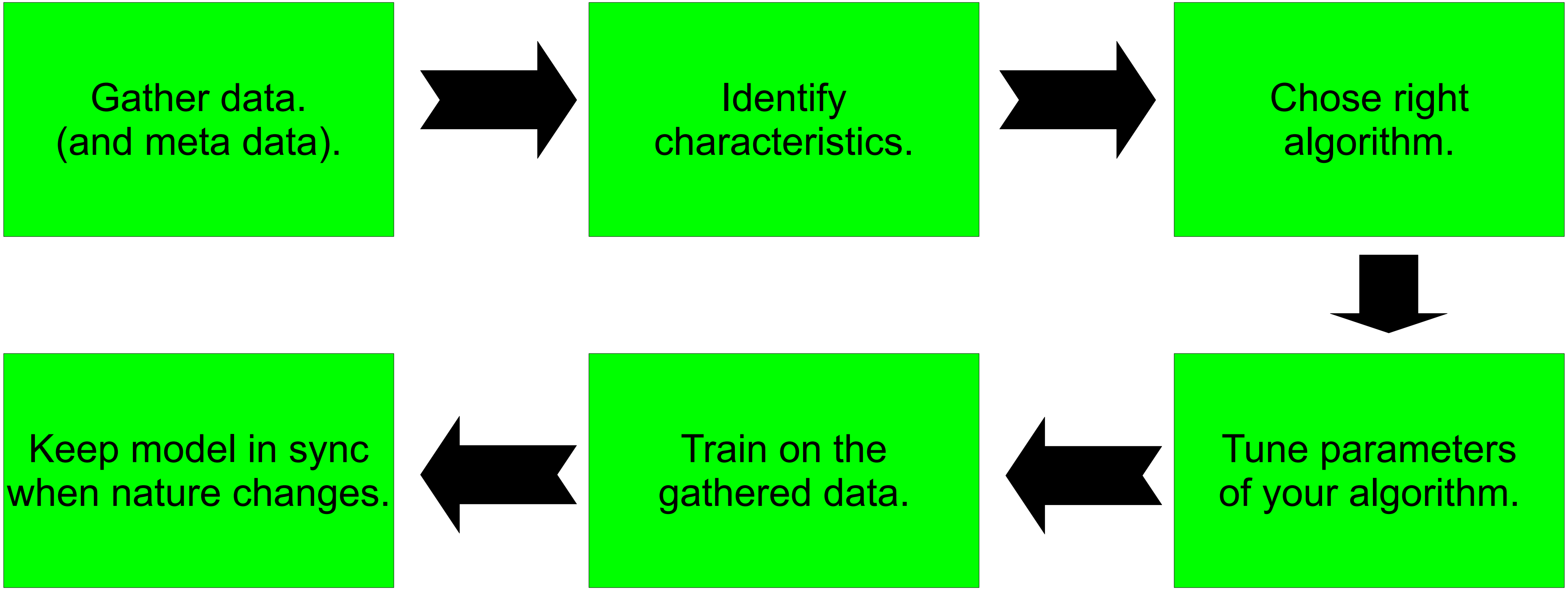
Train on the
gathered data.

Keep model in sync
when nature changes.

Nature changes?

- Spammers adapt to spam filters.
- Users write mails in different styles.
- Expand to new languages.
- ...

Machine learning pipeline



Gather data.
(and meta data).

Identify
characteristics.

Chose right
algorithm.

Tune parameters
of your algorithm.

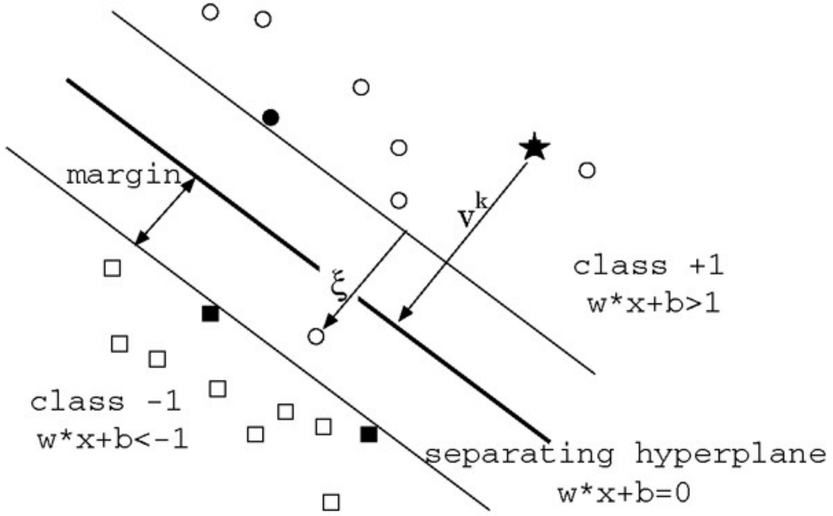
Train on the
gathered data.

Keep model in sync
when nature changes.

Introducing Mahout

Classification

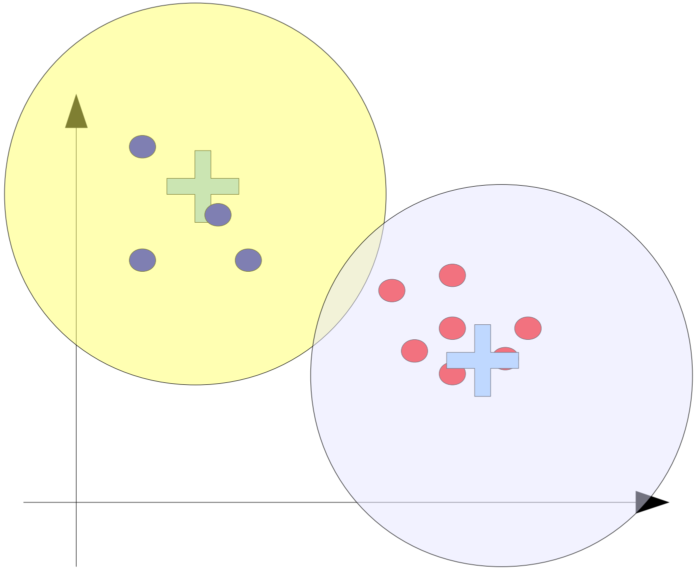
- Categorize data.
- Examples:
  - Identify spam mails.
  - Classify movies as “Action”, “Comedy” ...



Classification

- Naive bayes.
- Complementary naive bayes.
- Winnow/Perceptron
- Others upcoming.

Discovering groups of data



- Group data by similarity.
- Examples:
  - News articles by topic.
  - Developers by favorite modules.

Discovering groups of data

- Canopy.
- K-Means.
- Dirichlet based.
- PLSI.
- Others upcoming.

Recommendation mining

- Recommend items.
- Examples:
  - Find books a user my like.
  - Identify movies a user likes.

Upcoming

- More algorithms.
- More examples.

What Mahout can do for you

“Why should I participate?”

Jumpstart your project with proven code.

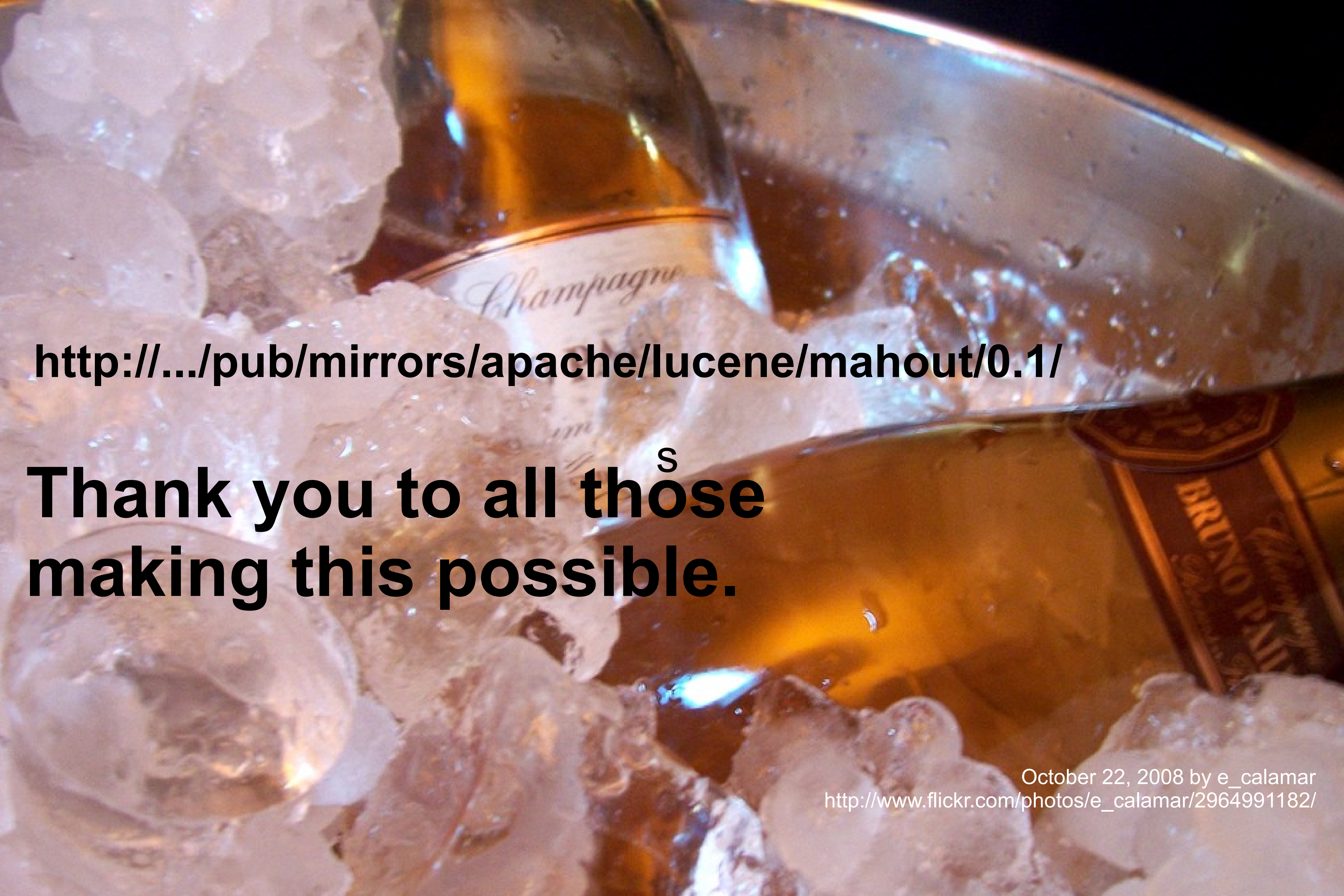
January 8, 2008 by dreizehn28
http://www.flickr.com/photos/1328/2176949559

Discuss with researchers and engineers.

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

Become a community member.

Doug Lea
Concurrent Programming in Java Second Edition
Design Principles and Patterns



http://.../pub/mirrors/apache/lucene/mahout/0.1/

Thank you to all tho^{s}se making this possible.

October 22, 2008 by e_calamar
http://www.flickr.com/photos/e_calamar/2964991182/

July 9, 2006 by trackrecord
http://www.flickr.com/photos/trackrecord/185514449

mahout-dev@apache.org

mahout-user@apache.org

- We need You:
  - Enthusiasm.
  - Mathematical knowledge.
  - Proficiency in Hadoop.
  - Interest in understanding data.

Some advertising

Berlin - June* at 5p.m.

newthinking store Berlin
Tucholskystr. 48

Hadoop** User/Developer Meeting Germany

* Exact date is set by speaker – that is you!

** Lucene, Tika, Solr, UIMA, Mahout, katta, ... people welcome.