Winter of Code – TU Berlin

Mahout Seminar

Isabel Drost

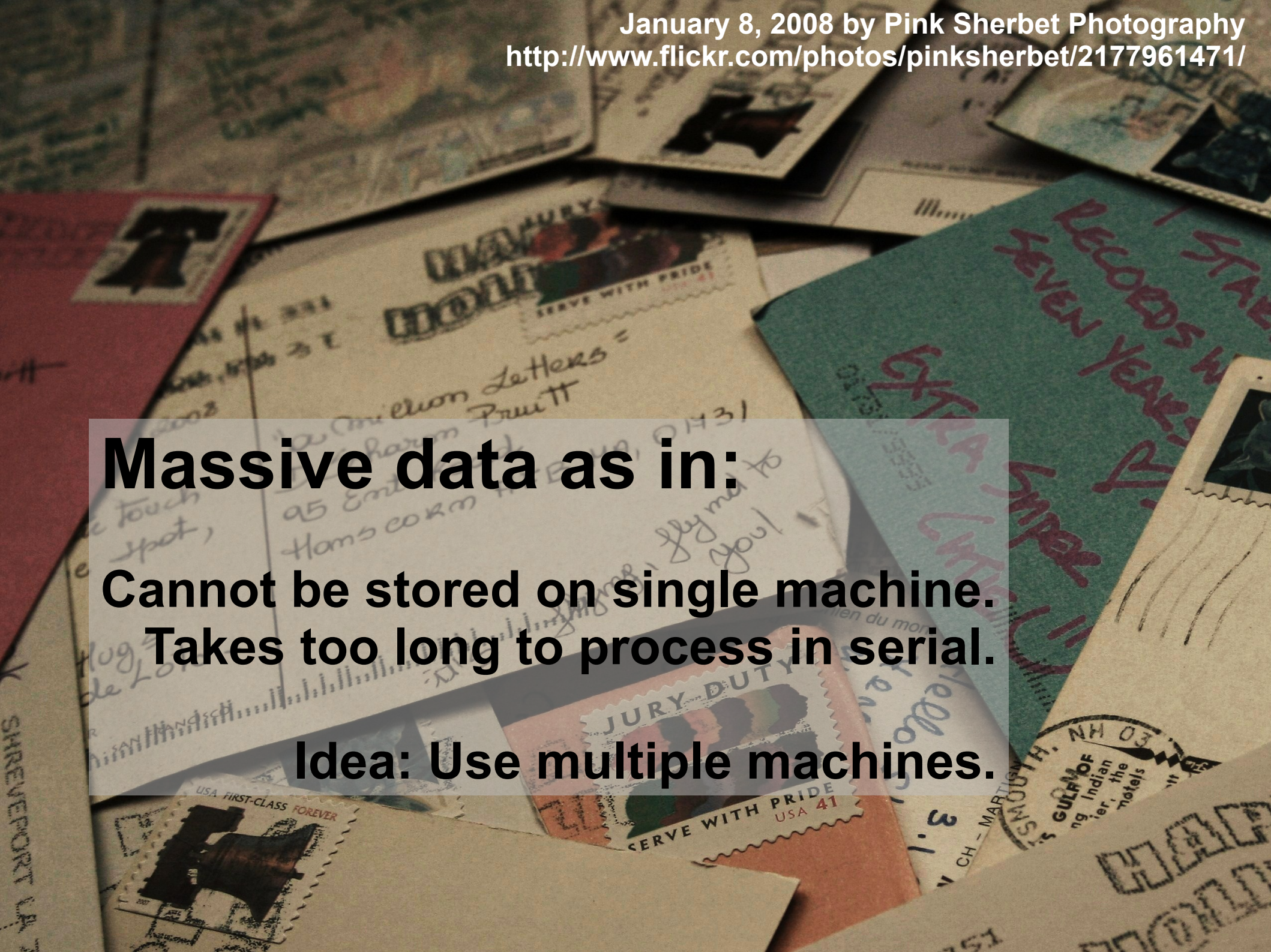
Nighttime:

Came to nutch in 2004.
Co-Founder Apache Mahout.
Organizer of Berlin Hadoop Get Together.

Daytime:

Software developer @ Berlin

Hello TU Students


January 8, 2008 by Pink Sherbet Photography
http://www.flickr.com/photos/pinksherbet/2177961471/
Massive data as in:
Cannot be stored on single machine.
Takes too long to process in serial.
Idea: Use multiple machines.

Presentation skills.

Choosing literature.

Citing sources.

Scientific texts.

Machine learning.

Schedule

- Choosing topics: 18. December 2009
- Literature list: 08. January 2009
- ToC: 15. January 2009
- Full paper: 29. January 2009
- Preliminary slides for review: 05. February 09.
- Block seminar: During term break.

Credits

http://github.com/MaineC/Playground/blob/master/projekt/credits

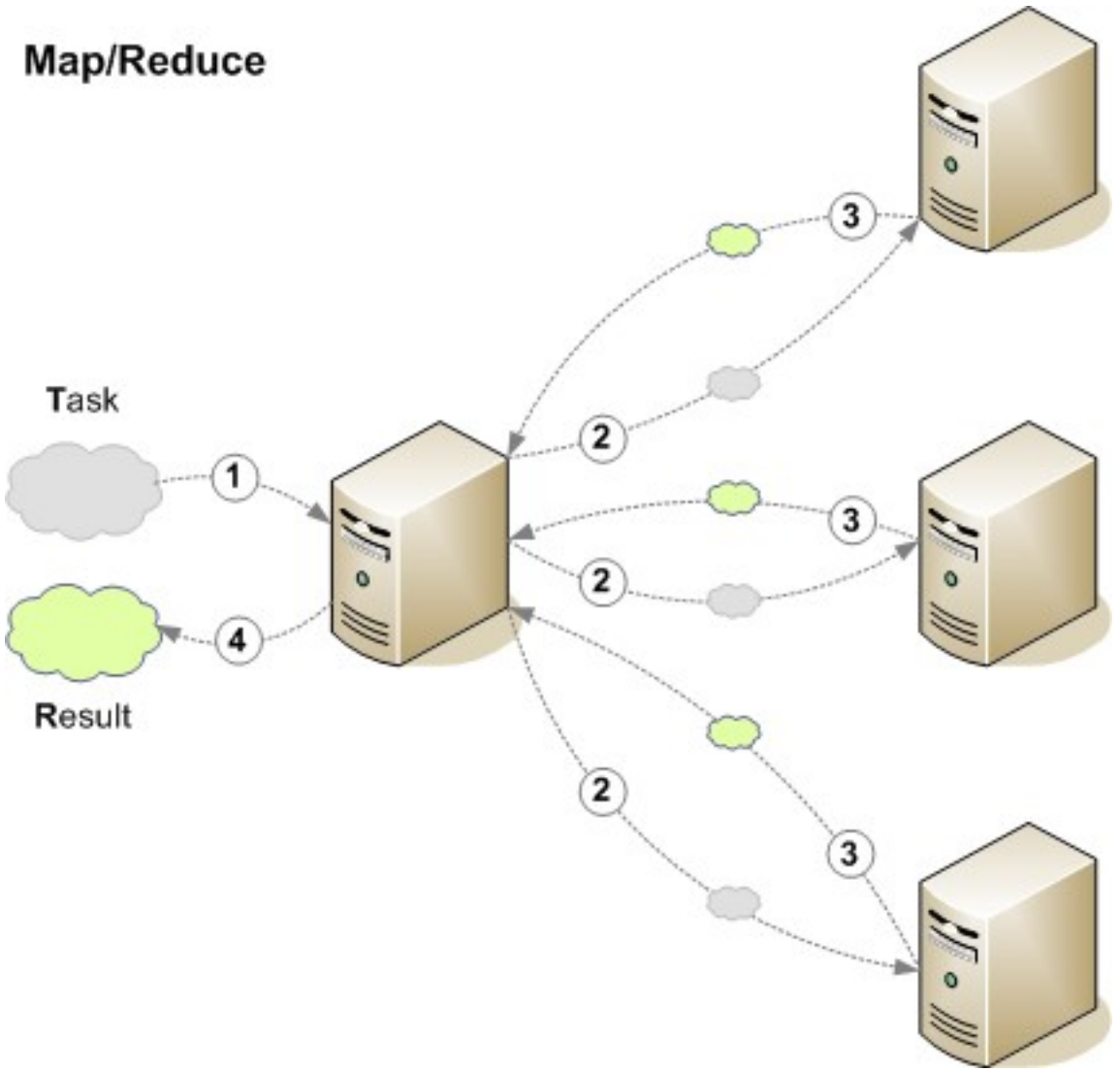
- 30 – clarity of paper.
- 30 – slides.
- 30 – presentation style.
- 30 – literature.

Total: 120
100 = 100%

Topics

Your proposals welcome.

Map Reduce and HDFS.



http://blogs.sun.com

NoSQL-/Document databases.

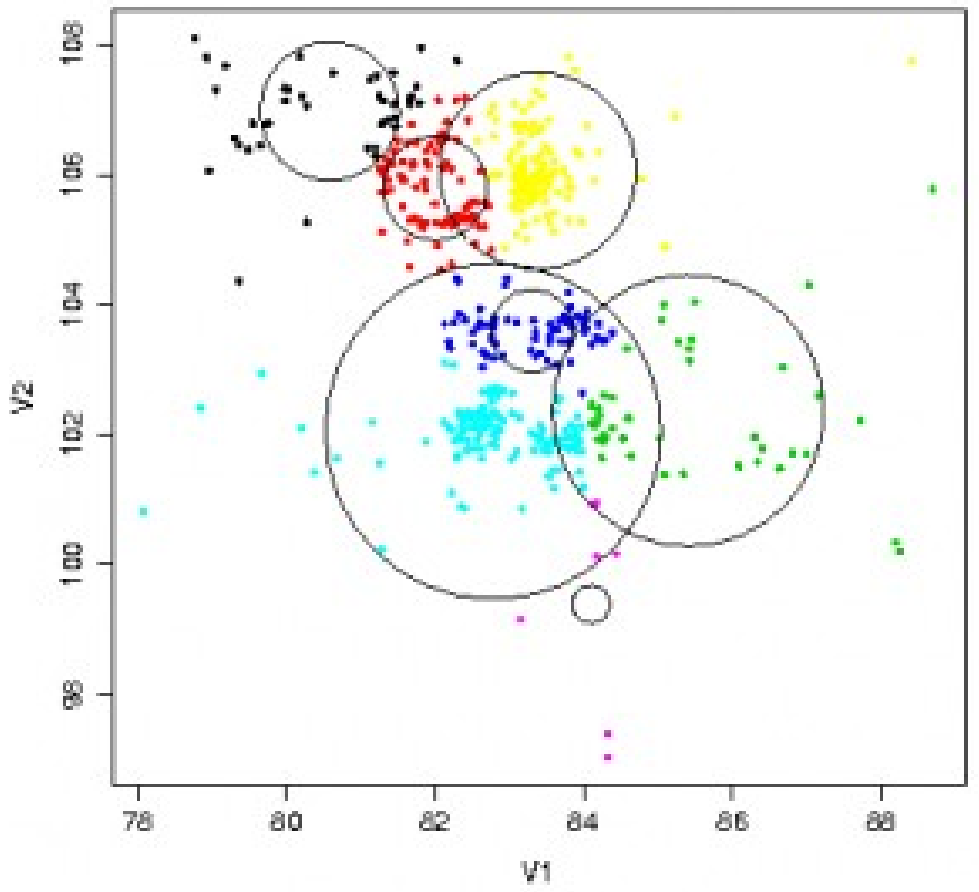
http://www.meetup.com/hbaseusergroup/ calendar/10950511/

(Near-) Duplicate detection.

http://www.flickr.com/photos/antichrist/122771132/

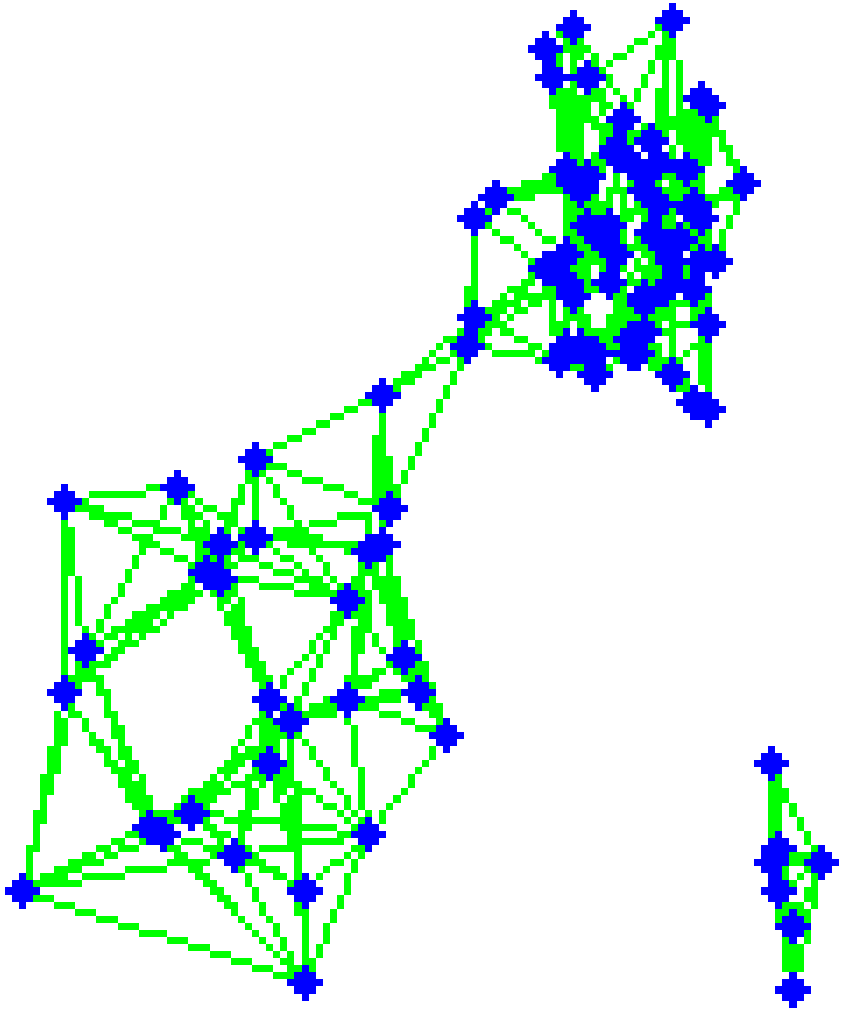
Topic tracking.

Document clustering: k-Means, dirichlet based clustering.



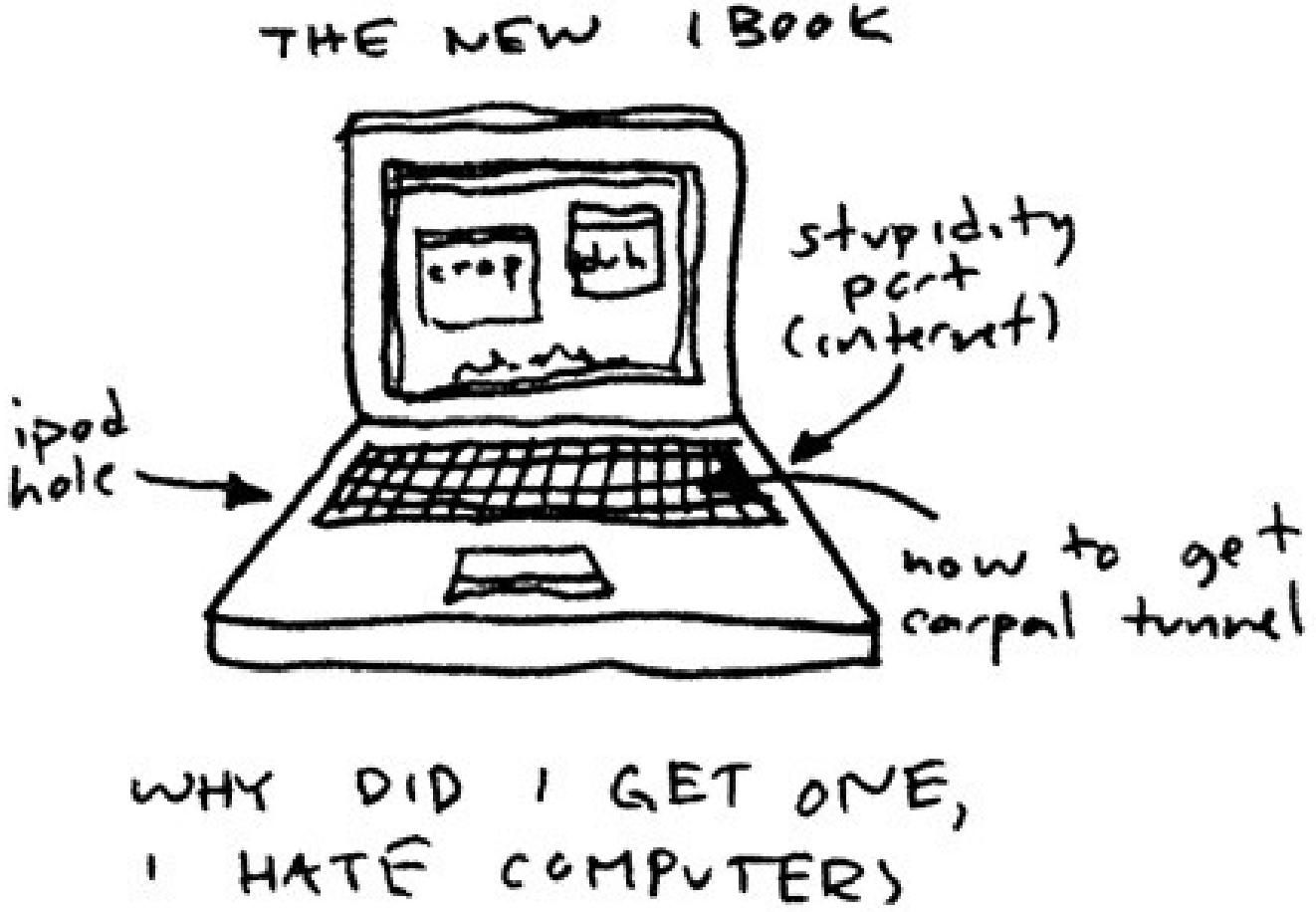
http://wiki.iosa.it/_media/spatial_analysis:kmeans1.png?w=300

Spectral clustering.



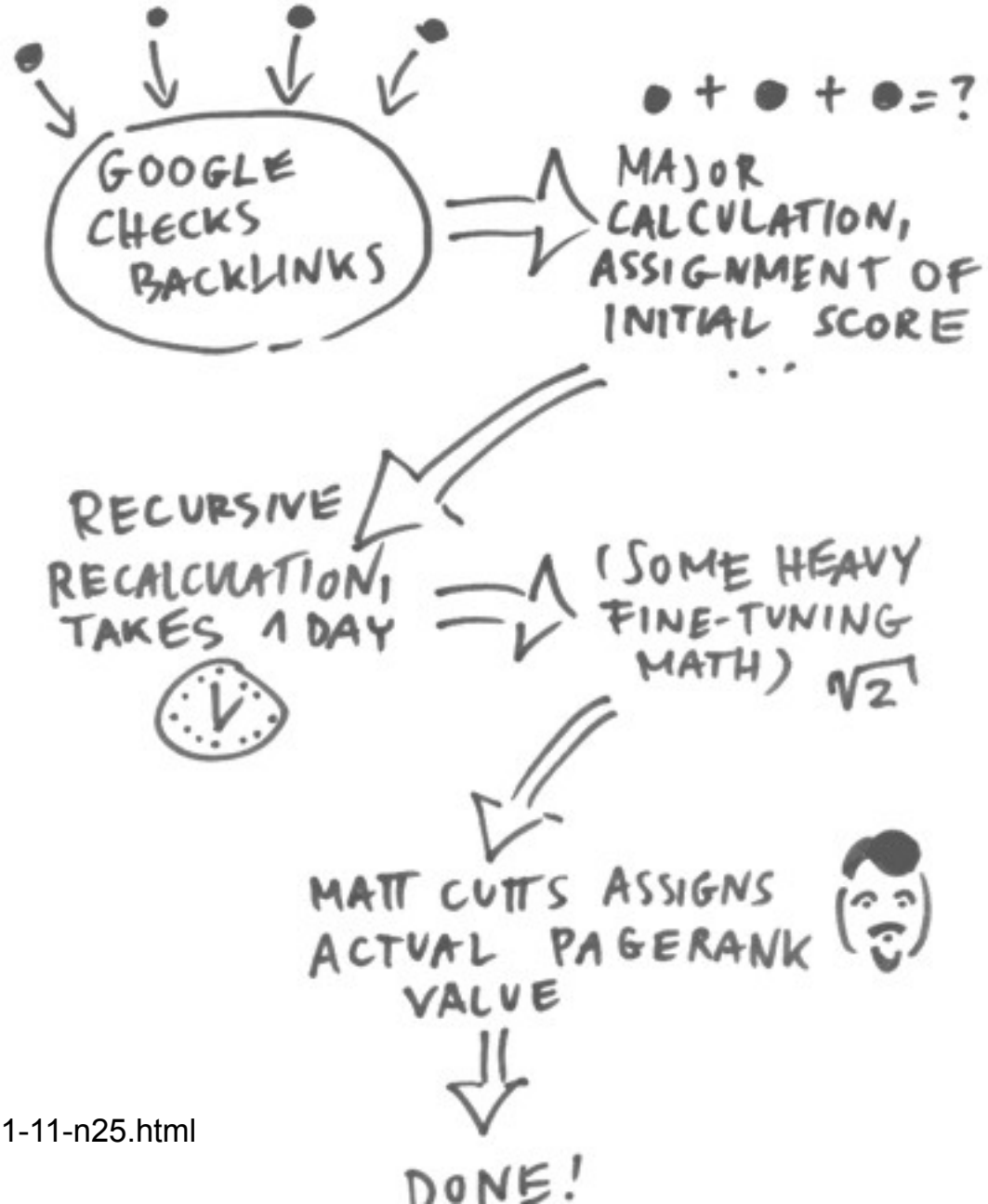
http://www.ml.uni-saarland.de/GraphDemo/DemoSpectralClustering.html

Opinion mining.



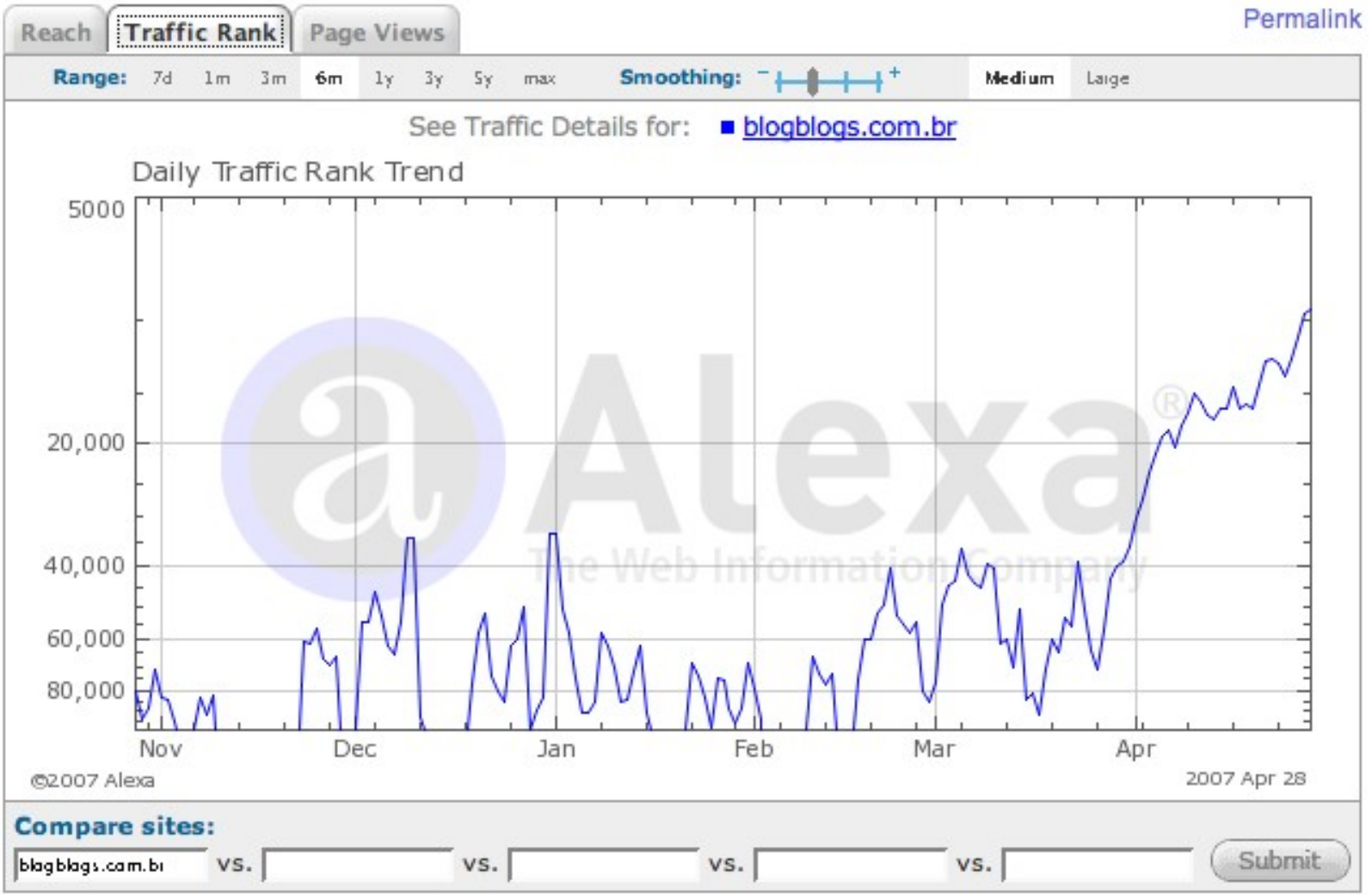
http://www.scienceforseo.com/opinion-mining/how-sentiment-analsysis-works/

Link based ranking of search hits.



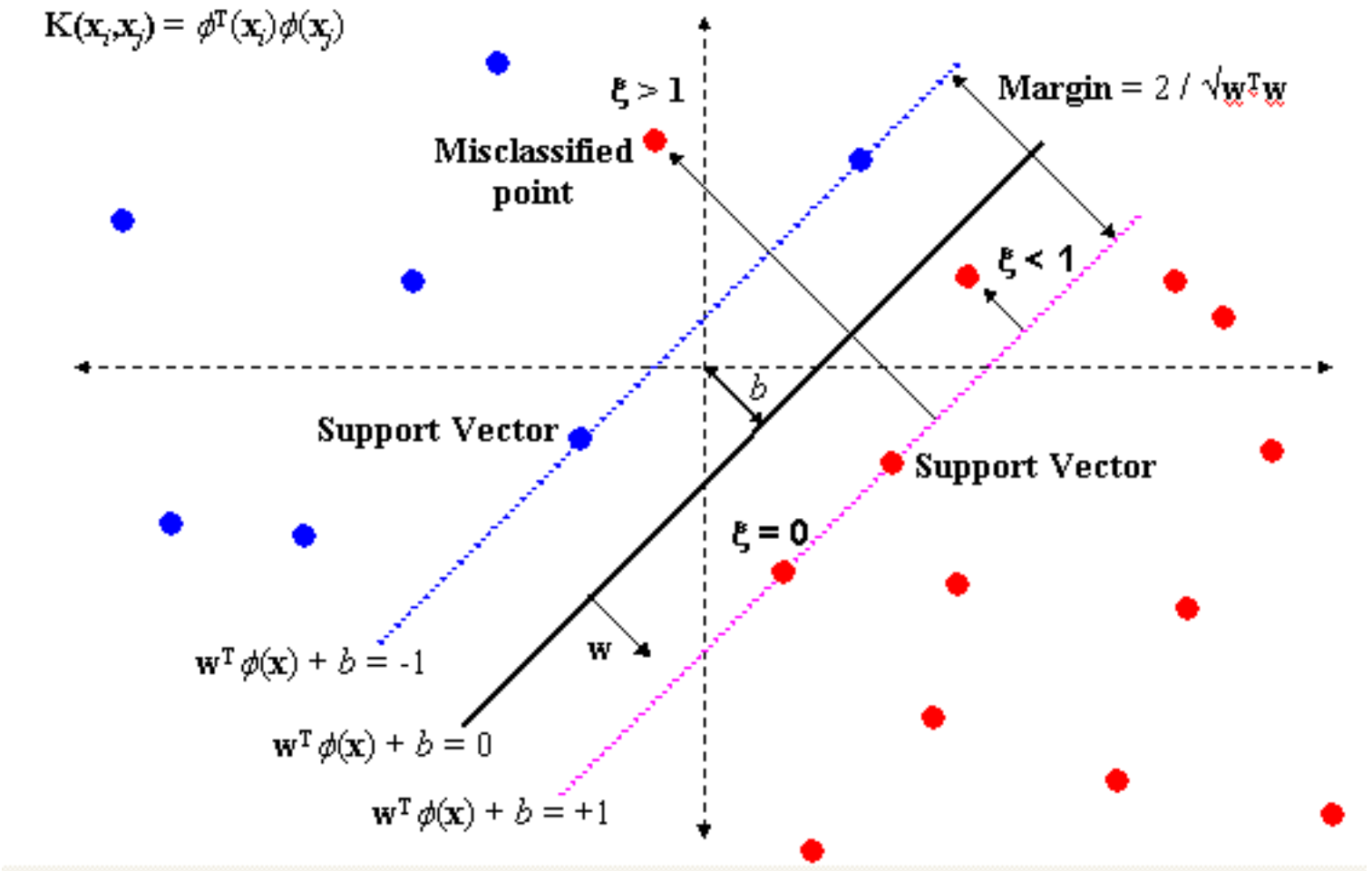
http://blogoscoped.com/archive/ 2007-01-11-n25.html

Learning Rankings



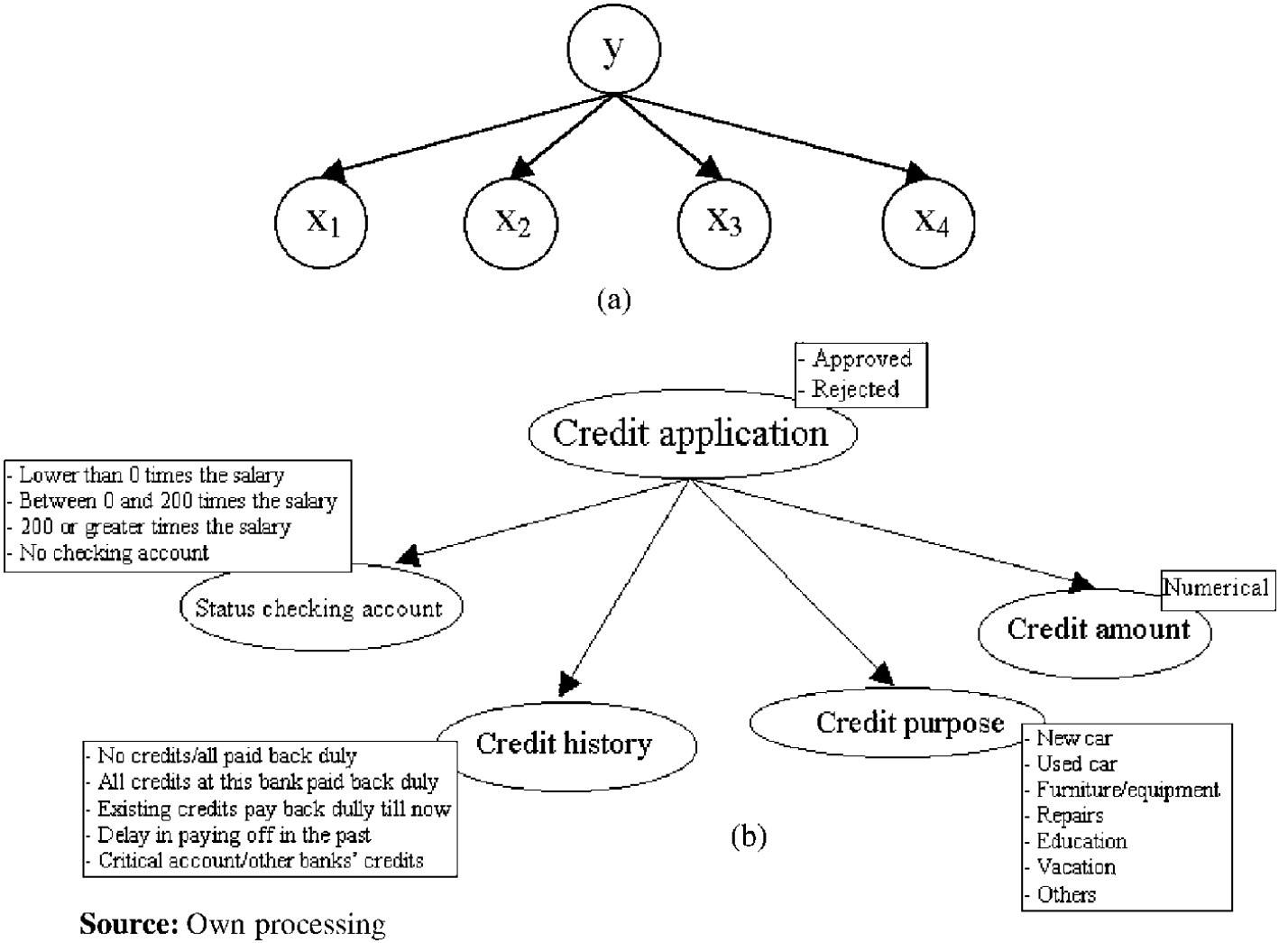
http://www.flickr.com/photos/mlemos/477874424/

Discriminative classification algorithms: Perceptron, Winnow, SVM



http://research.microsoft.com/en-us/um/people/manik/projects/trade-off/svm.html

Generative classification: Naive bayes, C-Bayes.



Source: Own processing

http://www.emeraldinsight.com/fig/0291060402002.png

HowTo – Ausarbeitung.

Literature

- Starting points online.
- Add your own:
  - Follow references.
  - Search for papers referencing publications.
- Search new publications:
  - Well known conferences (ICML, NIPS, KDD).
  - Scientific search engines.
- citeulike.org for organization.

ToC

- Title.
- ToC.
- Motivation.
- .
- Summary/ conclusion.
- Bibliography.

Style

- Referenced, numbered equations, illustrations.
- Objective, non-casual tone.
- Translation of established technical terms.
- Cited text not marked as citation: Plagiarism.

  L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998.

- LaTeX makes citing easy!
- Final format: Free format, e.g. Pdf.
- Send to isabel@apache.org or at github.