Machine Learning Open Source Software

(A PASCAL success story?)

Sören Sonnenburg

Fraunhofer FIRST.IDA, Berlin

joint work with

Cheng Soon Ong and Mikio Braun

FIRST

Fraunhofer Institut
Rechnerarchitektur
und Softwaretechnik

Outline

1. Introduction
2. History and Achievements
3. Future

What is Open Source Software?

Idea: Freedom to read, modify and redistribute source code

MS Windows network stack, MacOSX (BSD based)

TV (Sharp HDTV Aquos)

Mobile Phones (Motorola RAZR, Android)

Wireless routers (Linksys WRT)

Common: Free exchange of information, to avoid “reinventing the wheel”.

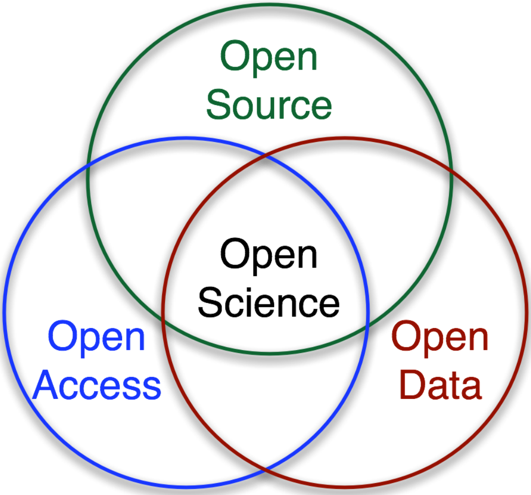
Open Source Definition (www.opensource.org)

- The Open Source Initiative (OSI) manages a license list of currently 65 approved open source licenses

Criteria to be open source:

1. Free redistribution
2. Must include source code
3. Derived works allowed
4. Integrity of the author's source code
5. No discrimination against persons or groups
6. No discrimination against fields of endeavor
7. License is redistributed
8. License must not be specific to a product
9. License must not restrict other software
10. License must be technology-neutral

Open Science



Open Source

Open Science

Open Access

Open Data

Open Access

> *Open access truly expands shared knowledge across scientific fields, it is the best path for accelerating multi-disciplinary breakthroughs in research.*
> *— Open letter to the U.S. Congress, signed by 25 Nobel laureates, (August 26, 2004)*

- Enabled by low-cost distribution on the Internet
- Open access literature is digital, online, free of charge, and free of most copyright and licensing restrictions. For example Creative Commons (`creativecommons.org`)
- Many journals (3096 according to `www.doaj.org`) have adopted the open access model (including JMLR, . . .)

Opening Machine Learning

- Hope to have a similar boost by adopting “open practices” in machine learning
  - *software and data* accompany paper
  - all openly licensed
- Some collections exists (UCI, Delve, Caltech, IDA Repository)
- How many machine learners publish software and data with their paper?
- Reasons? Misconception that open source renders commercial exploitation impossible?

Focus on Machine Learning Open Source Software

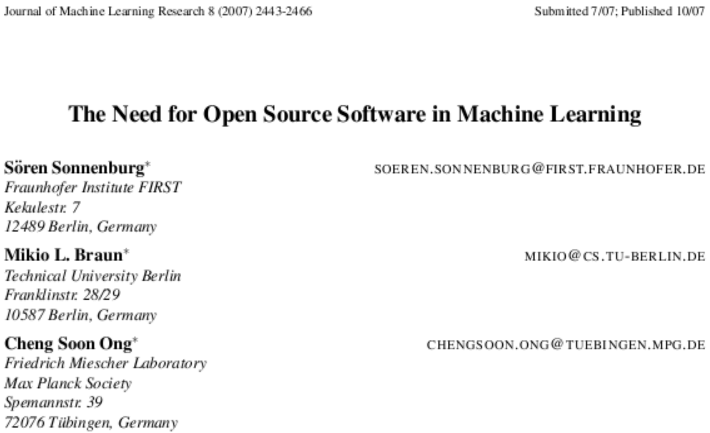
Advantages of Machine Learning Open Source Software

- **Reproducibility** of scientific results
- **Fair comparison** of algorithms
- **Problems uncovered quickly**
- Building on existing resources (rather than re-implementing)
- **Access to scientific** tools without restrictions
- Easier to combine different advances
- **Faster adoption** of ML methods in other disciplines and in industry
- Collaborative emergence of standards

Obstacles to an MLOSS community

- Publishing software is **not considered a Scientific contribution**
- **Misconception** — Opening the source **conflicts with commercial interests**
- The **incentive** for publishing open source software is **not high enough**
- Machine learning researchers may **not be good programmers**
- **Sloppiness hides problems** of newly proposed methods and eases acceptance at conferences and journals.
- **Tradition** — reviewers pass papers of similar quality

Read our Position Paper



Journal of Machine Learning Research 8 (2007) 2443-2466 Submitted 7/07; Published 10/07

The Need for Open Source Software in Machine Learning

Sören Sonnenburg* SOEREN.SONNENBURG@FIRST.FRAUNHOFER.DE
Fraunhofer Institute FIRST
Kekulestr. 7
12489 Berlin, Germany

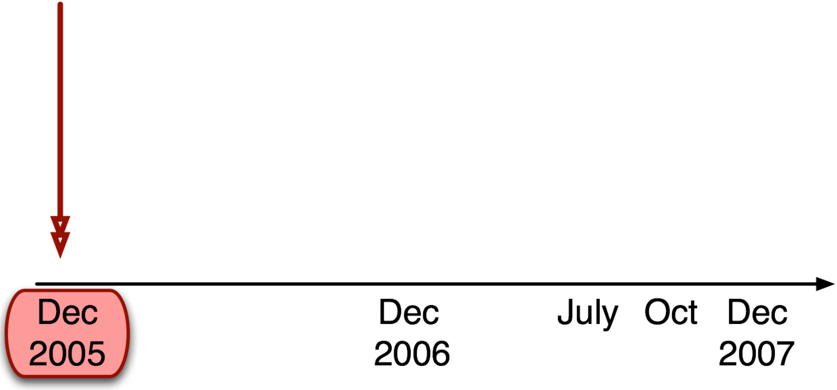
Mikio L. Braun* MIKIO@CS.TU-BERLIN.DE
Technical University Berlin
Franklinstr. 28/29
10587 Berlin, Germany

Cheng Soon Ong* CHENGSOON.ONG@TUEBINGEN.MPG.DE
Friedrich Miescher Laboratory
Max Planck Society
Spemannstr. 39
72076 Tübingen, Germany

. . . Bengio, Bottou, Holmes, LeCun, Müller, Pereira, Rasmussen, Rätsch, Schölkopf, Smola, Vincent, Weston, Williamson

Timeline

Machine Learning Tools Satellite Workshop



Dec 2005

Dec 2006

July

Oct

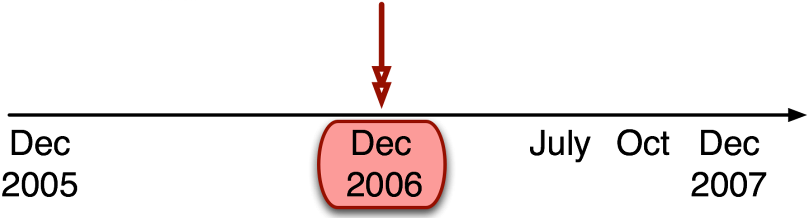
Dec 2007

Timeline

Machine Learning Tools Satellite Workshop

Workshop on Machine Learning Open Source Software

Neural Information
Processing Systems
Conference



Dec
2005

Dec
2006

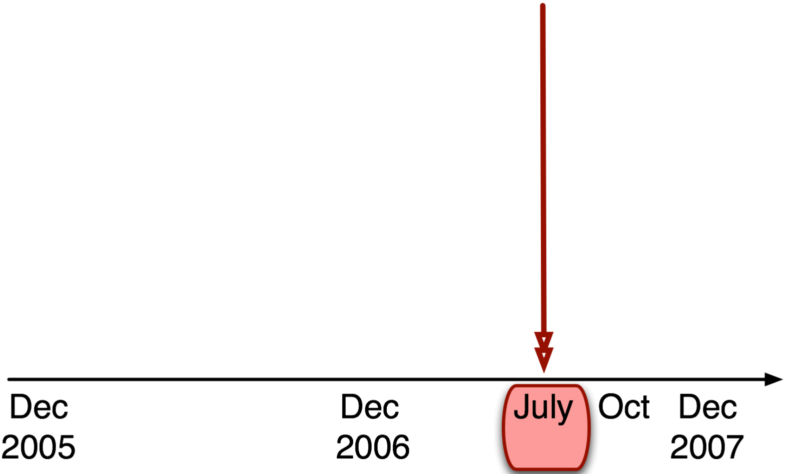
July

Oct

Dec
2007

Timeline

JMLR **Machine Learning Open Source Software**



Dec 2005

Dec 2006

July

Oct

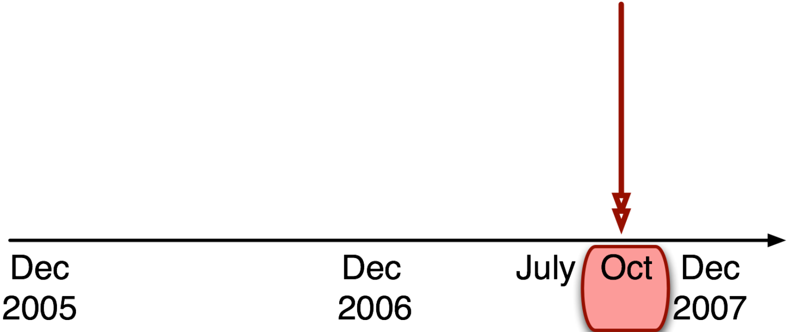
Dec 2007

Timeline

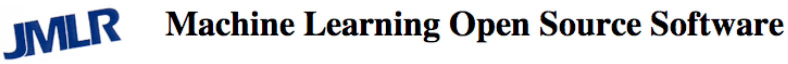
JMLR **Machine Learning Open Source Software**

Journal of Machine Learning Research 8 (2007) 2443-2466 Submitted 7/07; Published 10/07

The Need for Open Source Software in Machine Learning



Timeline

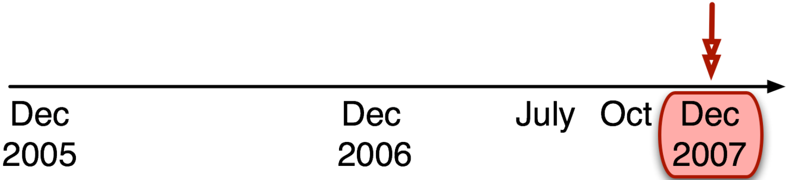


JMLR Machine Learning Open Source Software

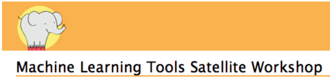
Journal of Machine Learning Research 8 (2007) 2443-2466 Submitted 7/07; Published 10/07

The Need for Open Source Software in Machine Learning

mloss.org
machine learning open source software



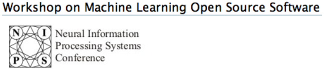
Workshops

- 

  Machine Learning Tools Satellite Workshop

  First PASCAL workshop

  - motivated by “The Mathworks” changing licenses. Affected people from Fraunhofer, Max-Planck, NICTA, INSA discussed and presented alternatives (octave, R, python,. . .)

- 

  Workshop on Machine Learning Open Source Software

  Neural Information Processing Systems Conference

  Second PASCAL/NIPS workshop

  - Open call for papers. Received 20 submissions, 8 accepted. 3 invited speakers (Weka, scipy, cvxopt)
  - Lively discussion, with the common themes:
    - incentives for researchers are missing
    - we should have a place to publish MLOSS
    - we still have a long way to go

New JMLR Track

JMLR **Machine Learning Open Source Software**

Contributions to http://jmlr.org/mloss/ should be related to

- Implementations of machine learning algorithms,
- Toolboxes,
- Languages for scientific computing

and should include

- A 4 page description,
- The code,
- A recognised open source license.

Community site mloss.org

All projects welcome

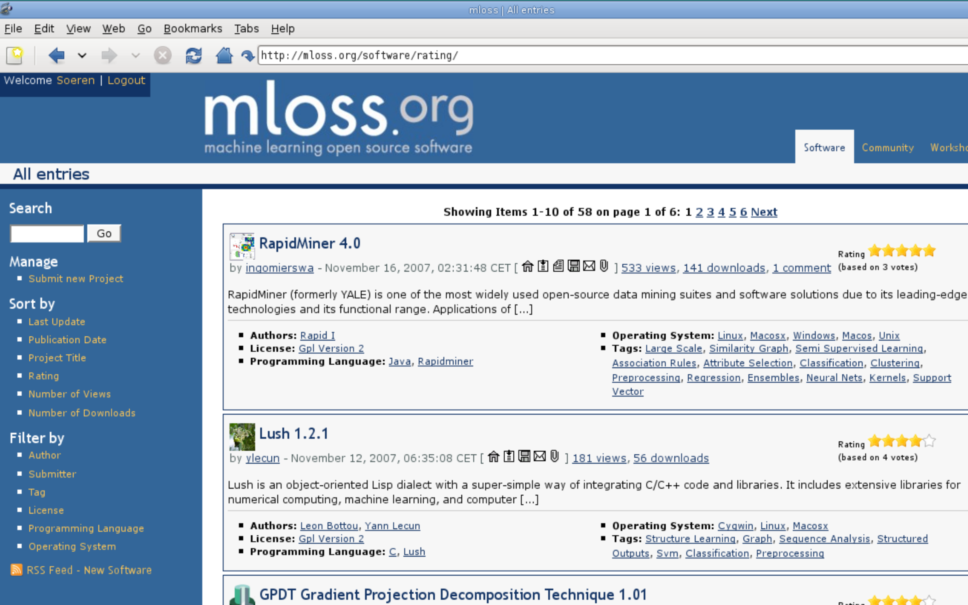
- Implementations of machine learning algorithms,
- Toolboxes,
- Languages for scientific computing
- Data readers, preprocessing
- Concrete applications

and should include

- A recognised open source license.
- Pointer to project homepage and download link

Contribute to http://mloss.org!

mloss.org Screenshot


mloss | All entries
File Edit View Web Go Bookmarks Tabs Help
http://mloss.org/software/rating/
Welcome Soeren | Logout
mloss.org
machine learning open source software
Software Community Worksho
Search
Go
Manage
Sort by
Filter by

All entries

Manage

- Submit new Project

Sort by

- Last Update
- Publication Date
- Project Title
- Rating
- Number of Views
- Number of Downloads

Filter by

- Author
- Submitter
- Tag
- License
- Programming Language
- Operating System

RSS Feed - New Software

Showing Items 1-10 of 58 on page 1 of 6: 1 2 3 4 5 6 Next

RapidMiner 4.0

by ingomierswa - November 16, 2007, 02:31:48 CET [⌂ 📋 📄 💾 ✉ 📎] 533 views, 141 downloads, 1 comment

Rating ★★★★★ (based on 3 votes)

RapidMiner (formerly YALE) is one of the most widely used open-source data mining suites and software solutions due to its leading-edge technologies and its functional range. Applications of [...]

- **Authors:** Rapid I
- **License:** Gpl Version 2
- **Programming Language:** Java, Rapidminer
- **Operating System:** Linux, Macosx, Windows, Macos, Unix
- **Tags:** Large Scale, Similarity Graph, Semi Supervised Learning, Association Rules, Attribute Selection, Classification, Clustering, Preprocessing, Regression, Ensembles, Neural Nets, Kernels, Support Vector

Lush 1.2.1

by ylecun - November 12, 2007, 06:35:08 CET [⌂ 📋 💾 ✉ 📎] 181 views, 56 downloads

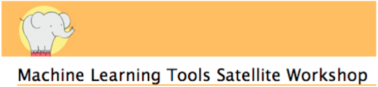
Rating ★★★★☆ (based on 4 votes)

Lush is an object-oriented Lisp dialect with a super-simple way of integrating C/C++ code and libraries. It includes extensive libraries for numerical computing, machine learning, and computer [...]

- **Authors:** Leon Bottou, Yann Lecun
- **License:** Gpl Version 2
- **Programming Language:** C, Lush
- **Operating System:** Cygwin, Linux, Macosx
- **Tags:** Structure Learning, Graph, Sequence Analysis, Structured Outputs, Svm, Classification, Preprocessing

GPDT Gradient Projection Decomposition Technique 1.01

The story so far...



Machine Learning Tools Satellite Workshop

Workshop on Machine Learning Open Source Software

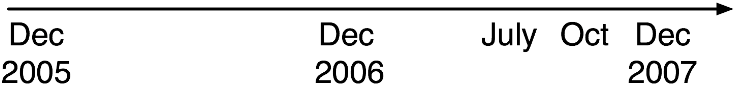
Neural Information Processing Systems Conference

JMLR Machine Learning Open Source Software

Journal of Machine Learning Research 8 (2007) 2443-2466 Submitted 7/07; Published 10/07

The Need for Open Source Software in Machine Learning

mloss.org
machine learning open source software



Dec 2005

Dec 2006

July

Oct

Dec 2007

Success Story? ⇒ Not yet

- JMLR track received 9 submissions, none accepted yet
  - Bigger established toolboxes are already published (e.g. Weka book)
  - It takes time to polish software projects to satisfy reviewers
  - New JMLR track not well known
- `mloss.org` currently has 180 registered users and 58 software projects
  - Collecting projects that are already out there
  - No collaborations / re-use of code yet
  - No lively discussion
  - Users mostly inactive

⇒ **How can we attract more (active) users?**

Call for help

mloss.org users needed

- Use software
- Rate software
- Comment on software
- Discuss in the forum how we can improve
- Discuss about data standards etc, etc.

Developers needed

- Help us to implement data standards
- Submit your software to mloss.org
- Help us to further develop and maintain the website.

Join the team!

We are open and need

- **Your Criticisms**
- **Your Ideas**
- **Your Feedback**
- **Your Contributions**

Plans for 2008

Events:

- See publications in JMLR-MLOSS
- Get lively discussions in `mloss.org`
- NIPS'08 workshop (if it gets accepted, otherwise NIPS satellite workshop)

Interoperability

- A common data exchange format

Data Set Standards

Goal: Develop a data exchange standard

- Currently, many data formats exist
  - ARFF
  - orange tab delimited
  - SVMlight, libsvm format
  - pyML, UCI
  - . . .
  - Post your thoughts!
    `http://mloss.org/community/standards/13/`
- A lot of time is wasted on converting data.

First Proposal: Use ARFF for dense vectorial data

- Used in Weka, code exists for R, matlab
- Subject to what the community thinks

Summary

Achievements

- Organized two PASCAL workshops on MLOSS
- Established a JMLR track for MLOSS (so far 9 submissions)
- Position paper on "The need for Open Source Software in Machine Learning"
- Community site `mloss.org` (58 projects, 180 users)

Future

- Data exchange standards
- Shall we (How can we) extend this approach to Open Data ?
- New ideas; Where should we go beyond 2008?