OpenML: a Networked Science Platform for Machine Learning (Abstract)

Joaquin Vanschoren J.VANSCHOREN@TUE.NL
M&CS, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Jan N. van Rijn J.N.VAN.RIJN@LIACS.LEIDENUNIV.NL
LIACS, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands

Bernd Bischl BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Giuseppe Casalicchio GIUSEPPE.CASALICCHIO@STAT.UNI-MUENCHEN.DE
Institut für Statistik, Ludwig-Maximilians-University Munich, Ludwigstrasse 33, D-80539 Munich, Germany

Michel Lang LANG@STATISTIK.TU-DORTMUND.DE
Fakultät Statistik, Technische Universität Dortmund, Vogelpothsweg 87, 44221 Dortmund, Germany

Matthias Feurer FEURERM@INFORMATIK.UNI-FREIBURG.DE
Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Georges-Köhler-Allee 052, 79110 Freiburg, Germany

OpenML is an online machine learning platform where researchers can *automatically* log and share data, code, and experiments in fine detail and organize them online to work and collaborate more effectively (Vanschoren et al., 2013). It is designed to create a *networked science* (Nielsen, 2012) ecosystem, allowing researchers all over the world to collaborate in large teams or completely in the open, while building on each others very latest ideas, data and results.

Data sets Data sets can be shared publicly or within *circles* of researchers. They can be uploaded or simply linked from existing scientific data repositories (e.g., mldata.org). For known data formats, OpenML will automatically analyze and annotate the data sets with measurable characteristics, so that they can be searched and analyzed based on this meta-data. Data sets can be repeatedly updated and are automatically versioned.

Tasks Data sets typically serve as input for scientific *tasks*. OpenML builds human and machine-readable descriptions of such tasks, defining which inputs are given, which outputs are expected to be returned, and what scientific protocols should be used. For instance, classification tasks will include performance estimation procedures (e.g. cross-validation), target features, evaluation measures, and require predictions or models as outputs. Tasks are similar to data mining challenges on platforms such as Kaggle (Carpenter, 2011), except that they are collaborative and real-time: others can immediately build on all shared results, while OpenML keeps track of who published what and when.

Flows Flows are implementations of machine learning workflows. They can be single algorithm implementations, scripts (e.g. in R) or workflows (e.g. in tools such as RapidMiner (van Rijn et al., 2013) and KNIME (Berthold et al., 2008)). They are again shared publicly or within *circles*, can be uploaded or linked from existing repositories (e.g. mloss.org), and updates are automatically versioned. Ideally, they are wrappers around existing software that take OpenML tasks as inputs. This allows automatic execution of algorithms on new data sets, but this is not required. Flows can also be annotated to facilitate search (e.g. to find algorithms able to deal with certain types of data).

Runs Runs are the results of executing flows on tasks, uploaded to the OpenML server. They are fully reproducible, containing details on the data set and flow versions, hyperparameter settings, and information on the authors and computational hardware. They are also reusable, containing non-aggregated results depending on the task. For instance, for classification this may include instance-level predictions, serialized models, runtimes, and user-defined evaluation measures. Where possible, runs are evaluated on the server to allow objective comparisons, using a broad range of evaluation measures and per-fold statistics. Runs are automatically linked to the underlying tasks, flows, and authors. This allows easy search, as well as direct comparisons across different data sets and flows.

Web services OpenML features an extensive REST API to allow easy communication with the server. Software tools can use this API to find and upload data sets, download tasks, find and upload flows, and download or upload runs. It also includes additional services such as authentication and tagging. All services are defined using XML schema, although we also offer JSON endpoints for accessing most data on the server.

Programming APIs Programming APIs are offered in Java, R and Python to allow easy integration into existing software tools. They make all OpenML services available as language-specific functions. For instance, using the OpenML[1] package for R, one can authenticate, search and download datasets, and upload the results of machine learning experiments in just a few lines of code.

Integration into machine learning toolkits OpenML is available as a WEKA (Hall et al., 2009) plugin through the WEKA package manager, allowing users to easily run WEKA algorithms on a large range of OpenML tasks, and immediately upload all results. We also offer plugins for MOA (Bifet et al., 2010), and HubMiner[2]. R is supported through the mlr[3] package. Other integrations, including RapidMiner (van Rijn et al., 2013), are currently being developed, and we hope to extend to other popular libraries soon.

Website OpenML.org is a website offering easy access to most OpenML functionality. It allows users to easily search and browse through all shared datasets, flows and runs. It compares all results obtained on specific tasks and flows, and has dedicated pages for each data set, task, flow and run with all known details. When logged in, you can also upload new data sets and flows, create new tasks, and organize information through tagging and wiki-like editing. It is also possible to comment on almost any shared resource. In the near future, it will also be possible to connect to and interact with other scientists through the website, and to organize work into *online studies* which can be linked (and backlinked from) to paper publications. Finally, the website offers an online guide with tutorials and developer documentation.

Open Source, Open Data OpenML is an open source project, hosted on GitHub[4] and the service is free to use under the CC-BY licence. When uploading new data sets and code, users can select under which licence they wish to share the data, and OpenML will clearly state these licences and citation requests online. OpenML has an active developer community, and everyone is welcome to help extend it, or post new suggestions through the website or through GitHub.

Activity OpenML currently contains close to 400 000 runs on about 1 200 data sets and 1 300 flows (including multiple versions). While still in beta development, it has over 300 registered users, over 1 000 frequent visitors, and the website is visited by around 100 unique visitors every day, from all over the world. It currently has server-side support for classification, regression, clustering, data stream classification, learning curve analysis, survival analysis, and machine learning challenges for classroom use.

Benefits for science Many sciences have made significant breakthroughs by adopting online tools that help organize, structure and analyze detailed scientific data online (Nielsen, 2012). Machine learning is a field where a more networked approach would be particularly valuable. Even today, a lot of research is only published in papers, in highly summarized forms such as tables, graphs and pseudo-code, which inhibits reuse and slows down research. Indeed, without prior experiments to build on, each study has to start from scratch, limiting the depth of studies and the ability to interpret and generalize their results (Hand, 2006; Keogh & Kasetty, 2003; Perlich et al., 2003). Moreover, it is often not even possible to rerun experiments because code, data, or experiment details are missing. This lack of reproducibility has been warned against repeatedly (Keogh & Kasetty, 2003; Sonnenburg et al., 2007; Pedersen, 2008), and has been highlighted as one of the most important challenges in data mining research (Hirsh, 2008).

OpenML alleviates these problems by allowing scientists to automatically share their experiments through the tools they are already using. OpenML integrations make sure that all necessary details are uploaded for future reference, ensuring reproducibility and interpretability. It also organizes all results online, allowing researchers to easily reuse and build on the results of others, thus enabling larger, more generalizable studies that were practically infeasible before.

New discoveries could by made simply by *querying* or *mining* all combined experiments to answer interesting questions. These question may have been nearly impossible to answer before, but are easily answered if a lot of data is already available. In addition, it becomes a routine part of research to answer questions such as “What is the effect of data set size on runtime?” or “How important is it to tune hyperparameter P?” With OpenML, we can answer these questions in minutes, instead of having to spend days setting up and running new experiments (Vanschoren et al., 2012). This means that more such questions will be asked, possibly leading to more discoveries.

Large-scale studies could be undertaken as a team, or hard questions could be tackled collaboratively, with many scientists contributing according to their specific skills, time

[1] https://github.com/openml/r

[2] http://mloss.org/software/view/574/

[3] https://github.com/berndbischl/mlr

[4] https://github.com/openml

or resources. Scientists from other domains can draw attention to an important scientific problem by adding new data sets and tasks. Machine learning researchers may suggest techniques, design custom-built workflows, run large-scale experiments, or improve code, while the scientists that contributed the data provide direct feedback on the practical utility of suggested approaches, interpret the generated models, and otherwise guide the collaboration to the desired outcome. Such collaborations can scale to any number of scientists. OpenML helps to coordinate the effort by organizing all results per task, so that everybody can track each other's progress, and discuss ideas and results online.

Benefits for scientists Individual scientists can also benefit from using OpenML. First, they gain *more time*. OpenML assists in most of the routine and tedious duties in running experiments: finding data sets, implementations, and prior results, setting up experiments, and organizing all experiments for further analysis. Moreover, when they share experiments, they can immediately compare them to the state of the art, and answer all kinds of routine research question in minutes by tapping into all shared data. Storing and organizing experiments online also means that they are available any place, any time.

Second, linking one's results to everybody else's has a large potential for *new discoveries*. It facilitates much larger, more convincing studies based on data by many other people, and scientists can interact with other minds on a global scale to answer questions, learn what others are doing, and forge new collaborations. Especially in large scientific studies, data management and sharing are very important, and OpenML offers a practical way to address these issues.

Third, OpenML helps scientists build their reputation by making their work more visible to a wider group of people. Scientists can indicate how they want people to cite their data, code and experiments, thus making frequent citations more likely. OpenML will also automatically track how often one's data, code and experiments are downloaded or reused in the experiments of others. When results are particularly good, they will appear more prominently in comparisons, and even when they are surprising (and possibly hard to publish) they can be shared and discussed online.

Community OpenML aims to engender an open ecosystem for machine learning research, both within and across scientific domains. We welcome scientists from all domains to post relevant data and collaborate online with the machine learning community to analyse and understand data better, together. More information and documentation can be found on the website (`http://openml.org`). As an open source platform, we warmly welcome all contributions, comments and suggestions.

References

Berthold, M, Cebron, N, Dill, F, Gabriel, TR, Kotter, T, Meini, T, Ohl, P, Sieb, C, Thiel, K, and Wiswedel, B. KNIME: The Konstanz information miner. *Studies in Classification, Data Analysis, and Knowledge Organization*, 5:319–326, 2008.

Bifet, A, Holmes, G, Kirkby, R, and Pfahringer, B. MOA: Massive Online Analysis. *Journal of Machine Learning Research*, 11:1601–1604, 2010.

Carpenter, J. May the best analyst win. *Science*, 331(6018): 698–699, 2011.

Hall, MA, Frank, E, Holmes, G, Pfahringer, B, Reutemann, P, and Witten, IH. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

Hand, D. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

Hirsh, H. Data mining research: Current status and future opportunities. *Stat.Ana.&Dat.Min.*, 1(2):104–107, 2008.

Keogh, E and Kasetty, S. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

Nielsen, Michael. *Reinventing discovery: the new era of networked science*. Princeton University Press, 2012.

Pedersen, T. Empiricism is not a matter of faith. *Computational Linguistics*, 34:465–470, 2008.

Perlich, C, Provost, F, and Simonoff, J. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, 4:211–255, 2003.

Sonnenburg, S, Braun, M, Ong, C, Bengio, S, Bottou, L, Holmes, G, LeCun, Y, Muller, KR, Pereira, F, Rasmussen, CE, Ratsch, G, Scholkopf, B, Smola, A, Vincent, P, Weston, J, and Williamson, R. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466, 2007.

van Rijn, J. N., Umaashankar, V., Fischer, S., Bischl, B., Torgo, L., Gao, B., Winter, P., Wiswedel, B., Berthold, M. R., and Vanschoren, J. A RapidMiner extension for open machine learning. In *RapidMiner Community Meeting and Conference 2013*, pp. 59–70, 2013.

Vanschoren, J., Blockeel, H., Pfahringer, B., and Holmes, G. Experiment databases. A new way to share, organize and learn from experiments. *Machine Learning*, 87(2): 127–158, 2012.

Vanschoren, Joaquin, van Rijn, Jan N., Bischl, Bernd, and Torgo, Luis. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.