

# Towards Scalable Online Machine Learning Collaborations with OpenML

Joaquin Vanschoren  
Eindhoven University of Technology  
j.vanschoren@tue.nl

## ABSTRACT

Is massively collaborative machine learning possible? Can we share and organize our collective knowledge of machine learning to solve ever more challenging problems? In a way, yes: as a community, we are already very successful at developing high-quality open-source machine learning libraries, thanks to frictionless collaboration platforms for software development. However, code is only one aspect. The answer is much less clear when we also consider the data that goes into these algorithms and the exact models that are produced. A tremendous amount of work and experience goes into the collection, cleaning, and preprocessing of data and the design, evaluation, and finetuning of models, yet very little of this is shared and organized in a way so that others can easily build on it.

Suppose one had a global platform for sharing machine learning datasets, models, and reproducible experiments in a frictionless way so that anybody could chip in at any time to share a good model, add or improve data, or suggest an idea. OpenML is an open-source initiative to create such a platform. It allows anyone to share datasets, machine learning pipelines, and full experiments, organizes all of it online with rich metadata, and enables anyone to reuse and build on them in novel and unexpected ways. All data is open and accessible through APIs, and it is readily integrated into popular machine learning tools to allow easy sharing of models and experiments. This openness also allows a budding ecosystem of automated processes to scale up machine learning further, such as discovering similar datasets, creating systematic benchmarks, or learning from all collected results how to build the best machine learning models and even automatically doing so for any new dataset. We welcome all of you to become a part of it.

## PVLDB Reference Format:

Joaquin Vanschoren. Towards Scalable Online Machine Learning Collaborations with OpenML. PVLDB, 14(13): 3418-3418, 2021.  
doi:10.14778/3484224.3484239

## BIOGRAPHY

Joaquin Vanschoren is an assistant professor at the Eindhoven University of Technology (TU/e). His research focuses on the automation of machine learning (AutoML) and meta-learning. He co-authored and co-edited the books “Automatic Machine: Methods, Systems, Challenges” and “Meta-Learning: Applications to

AutoML and Data Mining”, published over 100 articles on these topics, and received an Amazon Research Award, an Azure Research Award, the Dutch Data Prize, and an ECMLPKDD Demonstration Award.

He founded and leads OpenML.org, an open science platform for machine learning. He is a founding member of the European AI Associations ELLIS and CLAIRE, chairs the Open Machine Learning Foundation, and co-chairs the W3C Machine Learning Schema Community Group.

He has been tutorial speaker at NeurIPS and AAAI, and has given more than 20 invited talks, including ECDA, StatComp, IDEAL, and workshops at NeurIPS, ICML, and SIGMOD. He is datasets and benchmarks chair at NeurIPS 2021, program chair of Discovery Science 2018, general chair at LION 2016, demo chair at ECMLPKDD 2013, and he co-organizes the AutoML and Meta-Learning workshop series at NeurIPS and ICML from 2013 to 2021.

---

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing [info@vldb.org](mailto:info@vldb.org). Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 13 ISSN 2150-8097.  
doi:10.14778/3484224.3484239