# Machine Learning Software in Practice: Quo Vadis?

Szilárd Pafka
Epoch
2644 30th St 2nd Fl
Santa Monica, CA 90401
szilard.pafka@epoch.com

## ABSTRACT

Due to the hype in our industry in the last couple of years, there is a growing mismatch between software tools machine learning practitioners wish for, what they would truly need for their work, what's available (either commercially or open source) and what tool developers and researchers focus on. In this talk we will give a couple of examples of this mismatch. Several surveys and anecdotal evidence show that most practitioners work most of the time (at least in the modeling phase) with datasets that fit in the RAM of a single server, therefore distributed computing tools are very often overkill. Our benchmarks (available on github [1]) of the most widely used open source tools for binary classification (various implementations of algorithms such as linear methods, random forests, gradient boosted trees and neural networks) on such data show over 10x speed and over 10x RAM usage difference between various tools, with "big data" tools being the most inefficient. Significant performance gains have been obtained by those tools that incorporate various low-level (close to CPU and memory architecture) optimizations.

Nevertheless, we will show that even the best tools show degrading performance on the multi-socket servers featuring a high number of cores, systems that have become widely accessible more recently. Finally, while most of this talk is about performance, we will also argue that machine learning tools that feature high-level easy-to-use APIs provide increasing productivity for practitioners and therefore are preferable.

## CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**;

## KEYWORDS

binary classification; software implementations; training speed; memory footprint; accuracy

## REFERENCES

[1] Szilárd Pafka. 2016. A Benchmark for the Scalability, Speed and Accuracy of Machine Learning Libraries for Binary Classification. https://github.com/szilard/benchm-ml. (2016).


*KDD '17, August 13-17, 2017, Halifax, NS, Canada*

ACM ISBN 978-1-4503-4887-4/17/08.
https://doi.org/10.1145/3097983.3106683