Meet Them Where They Are: An Analysis of Visualization Use in ML Tutorials and Software Libraries

Ge Gao Yuxuan Xiong Dylan Cashman

Brandeis University

Motivation

- Visualization for Machine Learning (**VIS4ML)** research frequently focuses on individual use cases
- ML is growing so rapidly that there is an opportunity to reach a broader audience *as they are learning*
- By improving the visualizations found in educational ML resources, we can have broader impact with VIS Research
- We need to understand how to reach broader audiences by understanding the landscape of visualizations being used

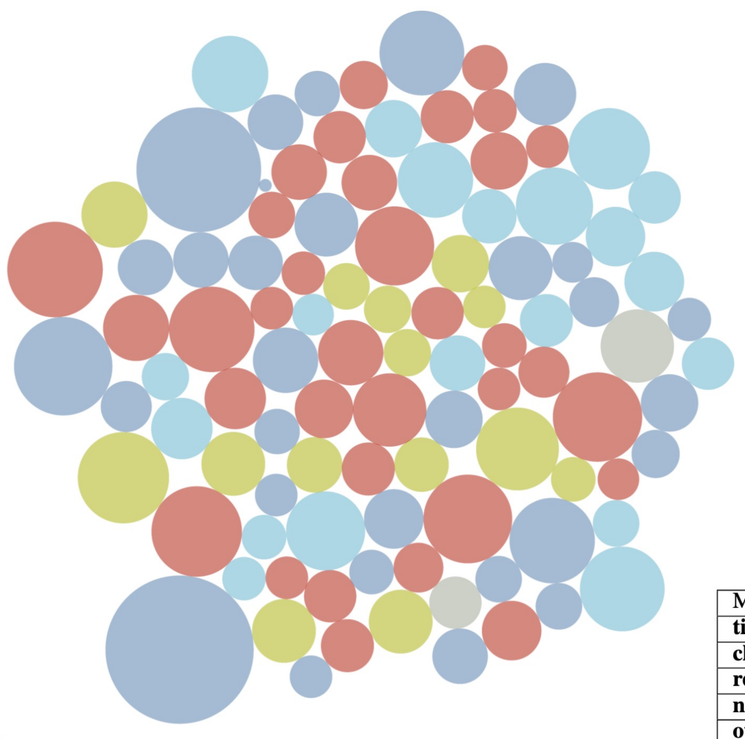
Method

We analyze the top 100 most-starred packages on Github tagged with Machine Learning

We focus on two types of popular repositories: **Tutorials** and **ML Packages**

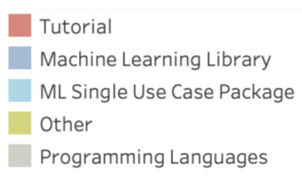
We analyzed all *visualizations* found in **tutorials** and all *merge requests* and *discussions* of visualizations in **ML libraries**

Results



Tutorials are very popular and frequently used to understand concepts, yet they use very simple visualizations.

ML Packages are frequently used to generate visualizations for these packages. Open source, unaffiliated packages may be easier to contribute to than those professionally managed.

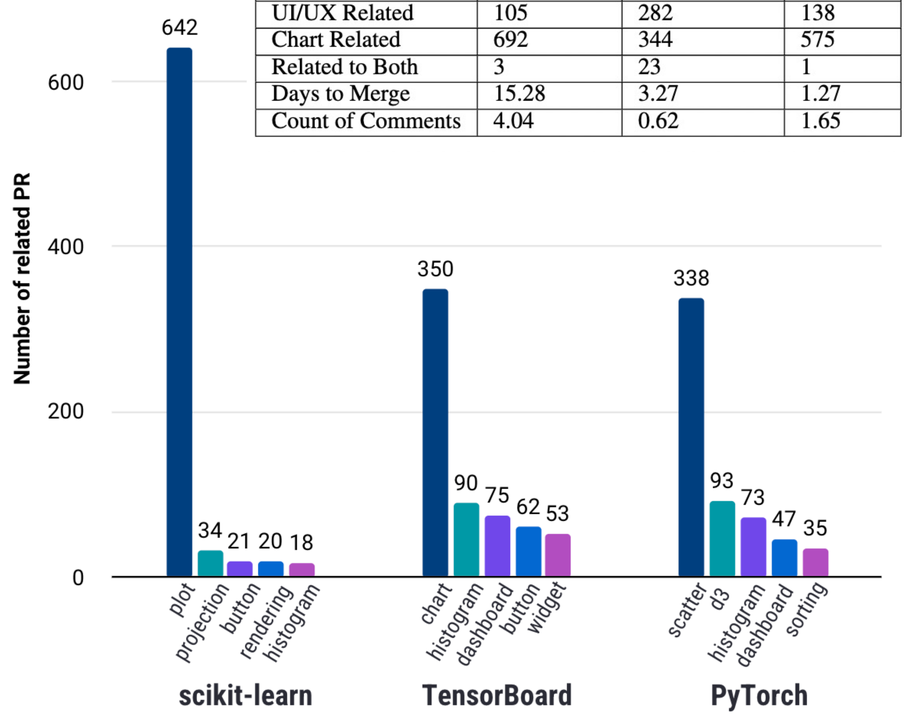


ML Task \Vis Type	scatter	line	node-link	other	total
time-series	1	11	0	1	13
clustering	16	0	0	0	16
regression	19	9	0	0	28
neural networks	8	6	6	5	25
other	19	21	5	12	57
total	63	47	11	18	139

Table 1: Classification of 139 visualizations found in 5 popular machine learning tutorials. Classifications of visualizations are based on Battle 2018 and classifications of ML tasks are based on tutorial content.

Table 2: Statistics of pull requests across three popular ML packages, as well as common visualization terms used in discussions in Github repositories

Metric	scikit-learn	TensorBoard	PyTorch
PR Count	794	603	712
Merged	632	541	37
UI/UX Related	105	282	138
Chart Related	692	344	575
Related to Both	3	23	1
Days to Merge	15.28	3.27	1.27
Count of Comments	4.04	0.62	1.65



Main Findings

- VIS4ML artifacts have great opportunity in reaching ML learners, improving their understanding of ML topics, and increasing the reach of the visualization community
- We recommend contributing well-constructed merge requests with examples and edge cases on popular ML data, with comparisons from existing capabilities

Future Work

Future research will evaluate existing VIS4ML research with the ML-learner audience of ML tutorials and ML packages

Contact Information

Ge Gao gegao@brandeis.edu
Yuxuan Xiong yuxuanxiong@g.ucla.edu
Dylan Cashman dylancashman@brandeis.edu

References

[1] L. Battle, P. Duan, Z. Miranda, D. Mukusheva, R. Chang, and M. Stonebraker. Beagle: Automated extraction and interpretation of visualizations from the web. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–8, 2018. 2

[2] L. Battle, D. Feng, and K. Webber. Exploring d3 implementation challenges on stack overflow. In 2022 IEEE visualization and visual analytics (VIS), pp. 1–5. IEEE, 2022. 1

[3] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. IEEE transactions on visualization and computer graphics, 23(1):61–70, 2016. 1