Table 1: An aspirational rubric for evaluating the rigor of ML practices.

Practice	**Good**	**Better**	**Best**
Problem definition	Have we clearly stated the ML task? Do we have a priori hypotheses? Do we know the predictions a domain expert would make manually?	Do we know the motivation for solving the problem? How much interpretability does the problem need?	Do we know our data? Do we know the confounding variables?
Model selection	Do we know the candidate algorithms for the ML problem?	Do we know our computational resources to fully train each model?	How much interpretibility does the problem need? How much each candidate algorithm can provide?
ML pipeline preparation	Do we have an held-out test dataset?	Have we tested our model on many different held-out datasets?	Have we tuned our model hyperparameters in cross-validation?
Hyperparameter selection	Do we know the different hyperparameters each model can use and why?	Did we use historically effective hyperparameters?	Did we search the full grid space and optimized our model?
Model evaluation	Have we chosen an appropriate metric to evaluate predictive performance?	Have we reported the predictive performance on a held-out test data?	Have we provided an average predictive performance of many model runs?
Model interpretation	Do we know if our model is interpretable?	If the model is not interpretable, do we know how to explain it? Have we checked for the effect of confounding variables?	Have we generated new hypotheses based on model interpretation to test model results?