Data vs Algorithms

If there is a need to make a choice between more or higher quality data vs a good algorithm, then the answer is always better data. The learning of a classifier or a regressor is always dependent upon the data. Better the data, better the performance of the classifier or the regressor.

There are several reasons for the need for a wide variety of algorithms (Wolpert 1997). Firstly, different algorithms work well with different types of data. For example, decision trees work better with categorical data, while neural network work better with numerical data. Secondly, the complexity of the data matters. With a nonlinear relationship between data, linear regression algorithm will not perform well as compared to other algorithms which can handle nonlinearities. The scalability and efficiency of data is another important consideration. With large datasets, algorithms like k-nearest algorithms can be computationally expensive, while logistic regression algorithms can work better. In addition, some algorithms work better for unsupervised learning while others work well for supervised learning, e.g., k-means algorithms vs support vector machines respectively. Some algorithms work better in handling missing data like decision trees, while others like support vector machines need more data pre-processing. Finally, some algorithms are naturally well suited to domain-specific problems, e.g., recurrent neural network for processing sequential data.

In other words, such insights are required for machine learning engineers to make judgement call on which algorithm to use based on the data and the problem statement.

Reference

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82.