# Machine Learning Glossary

As the hype around machine learning continues, we recognize the need to better define the most commonly used and sometimes mis-interpreted concepts. By clarifying these terms, we aim to help investors better understand the opportunities and challenges of applying Machine Learning in finance.

To start, we define **Machine Learning** as a collection of advanced models and algorithms for statistical prediction that can handle high dimensionality and nonlinearity.

**Artificial Intelligence (AI):** Artificial Intelligence is the science and engineering of making computers behave in ways that mimic human intelligence, such as learning, problem solving, and pattern recognition. Both Machine Learning and Deep Learning are subsets of Artificial Intelligence.

**Big Data:** Big data is an accumulation of both structured and unstructured data that is high in volume, velocity and variety. Big data is often difficult for traditional data processing applications to manage. Advances in computing power increase the accessibility of big data.

**Decision Tree:** Decision tree is a predictive algorithm that builds regression or classification models in the form of a tree structure. Finance researchers often apply tree models in portfolio selection process.

**Deep Learning:** A subset of Machine Learning, Deep Learning is a collection of algorithms that include specific approaches used for building and training neural networks. A model is "deep" if the input data passes through several levels of hierarchy before becoming output data.

**Natural Language Processing:** Natural Language Processing (NLP) is a subfield of Machine Learning for modeling human language. Financial institutions have applied NLP to extract important sentiments in digital news.

**Neural Network:** The Neural Network is one of the oldest statistical principles underlying Machine Learning. Neural networks are algorithms designed to learn from a complex set of observations and inputs in order to identify patterns.

**Out-of-sample prediction:** Out-of-sample prediction is commonly used to determine if a hypothesized predictor or model can accurately forecast a target variable. It is frequently used to evaluate the robustness of forecasting performance.

**Overfitting:** Overfitting occurs when a model fits too closely to a limited set of data points. An overfitted model captures too much of the noise in data, therefore making the model overly complex, which can then perform poorly out-of-sample.

**Random Forest:** Random forest is a popular method in Machine Learning that uses a group of independent decision trees to make an optimal prediction. The final prediction of the random forest is the average of the predictions of the trees.

**Signal-to-noise Ratio:** A measure of the true amount of predictability in a system. Machine Learning typically thrives in a high signal-to-noise ratio environment. Finance can be a noisy environment, making it more difficult to identify patterns.