WHITE PAPER

# Machine Learning and Big Data in Quantitative Investing

Uncovering Patterns in Financial and Alternative Data

MathWorks®

## Executive Summary

Machine learning enables computers or machines to learn from data directly without being explicitly programmed. By nature, machine learning models can capture nonlinearities better than traditional models can. To extract valuable information hidden in a large dataset, you need to use modern tools for processing big data and machine learning together.

What can you do with machine learning and big data in investing?

- Asset allocation and optimization
- Sentiment analysis with natural language processing
- Outlier and fraud detection
- Financial forecasting and price prediction

All these applications use machine learning in novel ways to solve fundamental investment challenges. However, successfully applying machine learning techniques requires data science skills—a special skill set with a low supply and high demand.

Addressing the data science skills gap, the August 2018 LinkedIn Workforce Report illustrated that the shortage of data scientists in the U.S. has grown to 151,717 people and already spread beyond the finance and tech industries.[1]

And then there's an opportunity to deploy machine learning models in traditional IT systems or on the cloud, where analytics work in real time for decision support and decision automation.

This white paper shows illustrates how you can apply machine learning and big data techniques to solve investment problems and improve investment performance.

## What Is Machine Learning?

*Machine learning* is a technique that enables computers or machines to learn from data directly without being explicitly programmed. In finance, machine learning provides an additional set of tools for developing predictive models that capture nonlinearities better than traditional models in use today.

Several taxonomies are known in the machine learning community. Generally, we can categorize machine learning by the types of problems we are trying to solve. For example:

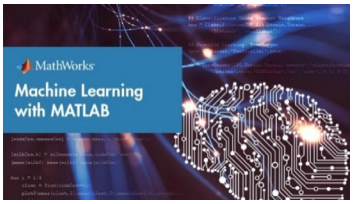
- **Supervised learning: learning from labeled data.** In supervised machine learning, the evidence comes in the form of data that's labeled with the correct (or desired) response. Supervised learning uses classification and regression techniques to develop predictive models.
- **Unsupervised learning: discovering patterns from unlabeled data.** In unsupervised machine learning, the evidence comes in the form of data without labeled responses. Clustering is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data.

[1] *https://news.linkedin.com/2018/8/linkedin-workforce-report-august-2018*



MathWorks®

- **Reinforcement learning: learning behaviors or actions.** The aim of reinforcement learning is to build a model that can perform a series of actions to maximize cumulative rewards. Instead of using a known set of input and output, reinforcement learning optimizes actions relative to a reward function. Fundamentally, reinforcement learning is like trial and error, in which the agent learns from positive and negative rewards based on its action.




**Learn More**

» *Machine Learning with MATLAB* - Ebook

## How Do You Decide Which Algorithm to Use?

Choosing the right algorithm can seem overwhelming—there are dozens of machine learning algorithms, and each takes a different approach to learning (Figure 1).

There is no single best method or one size fits all. Finding the right algorithm is partly trial and error—even experienced data scientists may have trouble telling whether an algorithm will work without trying it out. Among other factors, algorithm selection depends on the size and type of data you're working with and the insights you want to get from the data.

| SUPERVISED LEARNING | | UNSUPERVISED LEARNING | REINFORCEMENT LEARNING |
|---|---|---|---|
| CLASSIFICATION | REGRESSION | | |
| Logistic Regression | Linear Regression, GLM | K-Means, K-Medoids Fuzzy C-Means | SARSA[2] |
| Random Forests | Gaussian Process Regression | Hierarchical | Policy Gradient |
| Discriminant Analysis | Boosted Decision Trees | Gaussian Mixture | DDPG[3] |
| Naive Bayes | Decision Trees | Hidden Markov Model | Actor Critic |
| Nearest Neighbors | Support Vector Machines | DBSCAN[1] | Q-Learning |
| Boosted Decision Trees | Neural Networks and Deep Learning | Spectral Clustering | Neural Networks and Deep Learning |
| Random Subspace | | Neural Networks and Deep Learning | |
| Support Vector Machines | | | |
| Error-Correcting Output Code | | | |
| Neural Networks and Deep Learning | | | |

1. DBSCAN = Density-Based Spatial Clustering of Applications with Noise
2. SARSA = State–action–reward–state–action
3. DDPG = Deep Deterministic Policy Gradient

*Figure 1. Common machine learning algorithms.*


**Learn More**

» *Mastering Machine Learning: A Step-by-Step Guide with MATLAB* - Ebook

## What Is Big Data?

*Big data* generally refers to a large volume of data that is hard to process using existing techniques that require in-memory computation. Data used in or stored by financial institutions that is typically considered big data includes:

- Historical tick data of more than 1000 securities traded in the past 10 years
- Billions of credit card transactions
- News data related to all securities in the past 10 years
- Telephone communications of 1000 client-facing employees

It is worth noting that a computational algorithm for calculating mean and standard deviation by group when the data does not fit in memory can be much more complicated than when the data is small enough to fit in memory.

To extract valuable information hidden in big data, it is necessary to employ modern tools for processing big data and applying machine learning to find patterns in the data.

### Bigger Data, More Insights, Better Answers

The most basic example of the role of data in finance is technical analysis in which the relationships among price, volume, and time are aggregated into technical indicators for predicting future movements in price.

As the market has become more efficient, we can see the declining benefits of using technical indicators, especially on the most observable data like end-of-day data. Instead of using end-of-day data alone, investors have integrated shorter timeframe data like intraday and tick data into investment analysis. In addition, investors have enhanced investment performance using many different types of data such as news, social media posts, satellite images, and point-of-sale data. The common characteristics of this data are that the datasets are large and they're not simple to process in a timely manner.

### Accessing Big Data at a Glance

Numeric data is the most frequently used data type in finance, followed by text data. Although there are some use cases of imaging data (e.g., satellite images), it is not widely accessible for use in finance.

The first challenge in working with big data is determining how to access large datasets, as they come in many different forms and are stored in various types of systems:

- **Files:** Many datasets consist of a large number of small and medium-sized files. The number of files can grow quickly, and the files often do not fit into the memory of a single computer. These files typically reside within one or more directories on a shared drive and may consist of delimited text, spreadsheets, images, videos, and various proprietary formats.
- **Databases:** A wide range of database types are used to store and manage big sets of data in finance, including relational, graph, and document databases.
- **Hadoop:** Hadoop® is a system for storing and processing big datasets based on distributed computing and storage principles. It comprises two major subsystems that coexist on a cluster of computer servers (Figure 2):
  - Hadoop Distributed File System (HDFS): A large failure-resistant file system
  - YARN: An application scheduling framework that manages applications that run on Hadoop, including batch processing frameworks such as MapReduce and Spark™, and SQL interfaces such as Hive and Impala

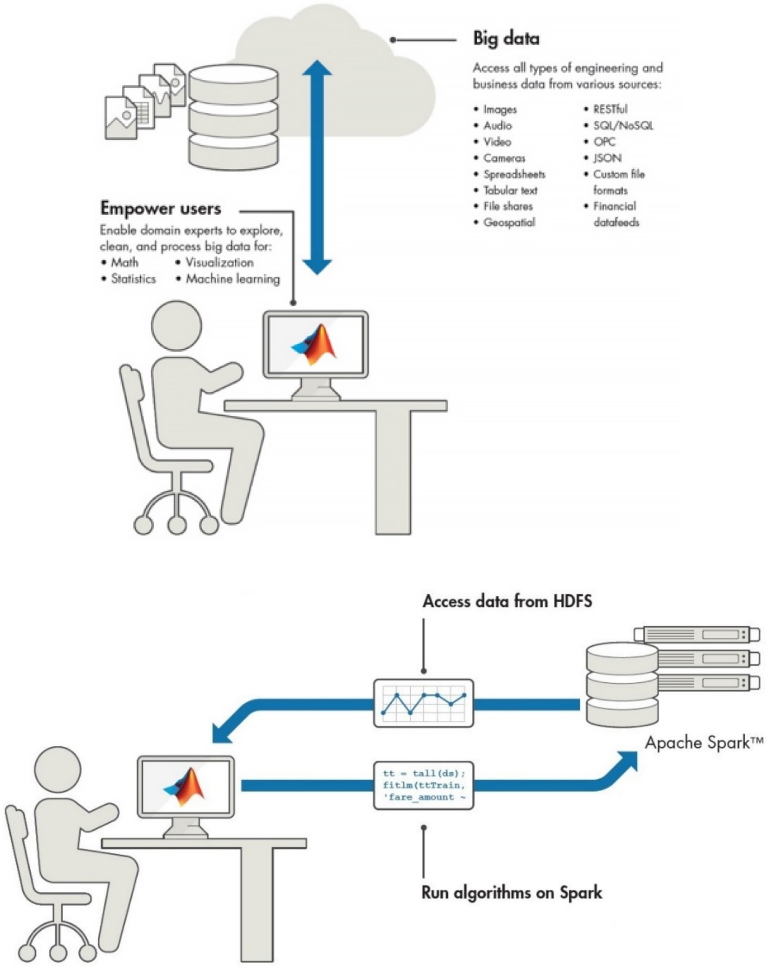



*Figure 2. Using MATLAB with data in HDFS and on Spark.*

## What Can You Do with Machine Learning and Big Data in Investing?

Consider using machine learning and big data when you have a complex task or problem involving a large amount of data and lots of variables, but no existing formula or equation. For example, machine learning and big data techniques are a good option when:

- The nature of data is unstructured (e.g., a combination of text, image, audio, or video).
- You need to quickly respond to large amounts of or high-velocity data, as in trade execution.
- Expert knowledge, handwritten rules, and equations are too complex to model, as in news sentiment analysis.
- The nature of the data keeps changing, and the program needs to adapt, as in asset allocation (Figure 3), automated trading, energy demand forecasting, and price trend prediction.

» *Asset Allocation - Hierarchical Risk Parity (Code Example)*

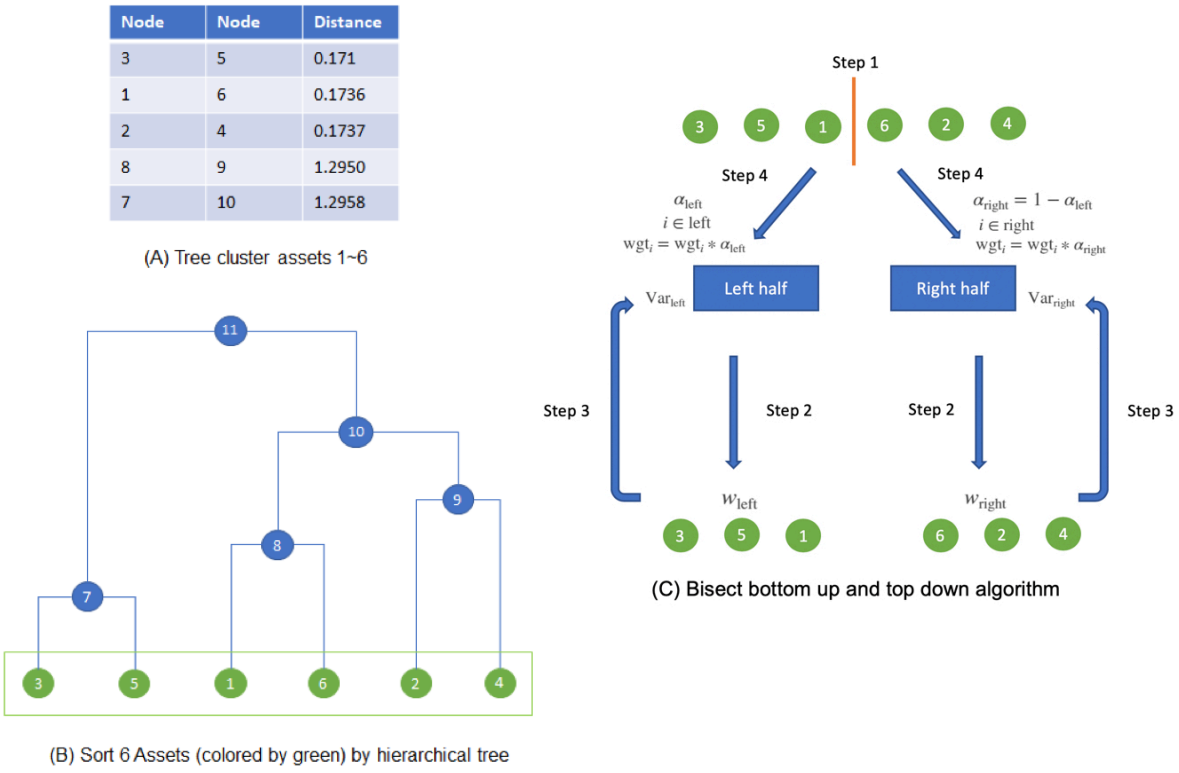



*Figure 3. Asset allocation workflow using hierarchical risk parity analysis (HRP) in MATLAB.*

## Is Machine Learning More Effective at Investing than a Human?

Although investment teams have increasingly adopted machine learning for investing and big data, the main usage is frequently in quantitative investment groups due to the computational nature of machine learning on large datasets. The main application of machine learning in portfolio management is to identify trading signals or create trading indicators on price movements from a large amount of data. In fact, with news trends, social media data, or satellite images, the amount of data has been increasing while the required reaction time can be as fast as a millisecond or even a microsecond. As a result, the availability of such big data rapidly drives the demand for investment teams to draw statistical inference from the data and create a predictive model using machine learning.

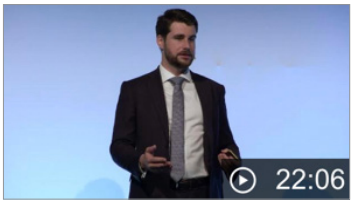
However, there are still areas in which humans typically outperform machine predictions, such as anticipating long-term trends or regime changes. Human analysts will continue to play a critical role.

### Which Techniques Are Most Effective?

Table 1 provides a comparison of the effectiveness of fundamental analysis and quantitative methods (both analyst-driven) and a machine learning and big data approach.

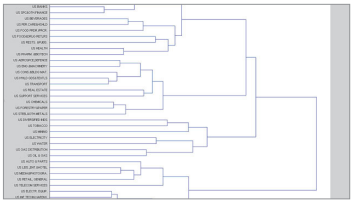
| Is the technique effective for: | Fundamental Analysis | Traditional Quantitative Methods | Machine Learning and Big Data Techniques |
|---|---|---|---|
| Predictivity based on time horizon | | | |
| Long-term trend | Yes | No | No |
| Short-term trend | No | Yes | Yes |
| Intraday trend (high-frequency data) | No | Maybe | Yes |
| Predictivity based on data type | | | |
| Structured data | Yes | Yes | Yes |
| Unstructured data | No | No | Yes |
| Small amount of data | Yes | Yes | No |
| Large amount of data | No | No | Yes |
| Ease of model interpretation | Easy | Moderate | Difficult |

## Machine Learning Examples




*Aberdeen Standard on Asset Allocation, Machine Learning, and High-Performance Computing* - Video

See how Aberdeen uses machine learning to analyse financial market trends and power innovative asset allocation strategies.



*Banque Cantonale Vaudoise Speeds Financial Analysis Tasks* - User Story

Learn how BCV Asset Management models industrial indices using a clustering model.

## Migration from Quants to Data Scientists

Who will do the analysis and build the analytics? To successfully create machine learning models, you need people who possess three types of knowledge:

- **Machine learning:** They understand how to apply statistical analysis and machine learning techniques to solve problems.
- **Computing:** They know the basics of coding, data management, and computing infrastructure.
- **Domain expertise:** They need to evolve to understand the financial modeling behind their projects.

Data scientists usually have a combined knowledge of machine learning and computing but may not have the domain expertise. It is very rare to find people who have all three, which makes it difficult to staff due to the intense competition for these skills. Almost every week articles are written about the shortage of data scientists, and it is recognized as a global issue.

In the financial services industry, quantitative analysts play an integral role in financial modeling. Compared with the required knowledge for data scientists, quants already have domain expertise in finance as well as computing skills. The only thing that they need to add is machine learning knowledge.

Rather than competing to hire data scientists who may not have the appropriate domain expertise, many financial institutions have hired and trained quants to do data science work. They can quickly learn how machine learning works and apply these techniques to solve the problem in finance.

## Machine Learning and Big Data with MATLAB

Using MATLAB®, teams can try out more ideas, which leads to more effective and efficient results in shorter time frames than with alternatives (e.g., Python® and R).

To maximize efficiency, focus your team's efforts on these four steps:

**1. Access and Explore Big Data**

MATLAB is designed to handle large amounts of data and various data types, including numeric, text, news, and social media data.

*"Our previous system was so tedious and our datasets are so large that I don't think this would have been possible without MATLAB and its ability to handle big data and interact directly with Bloomberg and our database."*

**— Ananthi Jegan, Olam CFSG**

**2. Preprocess Big Data**

MATLAB provides tools that make it easy to do high-speed processing of large datasets. In addition, its numeric routines scale directly to parallel processing on clusters and the cloud.

*"My expertise is in finance, not programming. To perform sophisticated analysis on vast amounts of data, I needed software that was easy to use and included many of the functions I needed. With MATLAB I can do everything in one environment, and that is a real benefit."*

**— Omid Rezania, CalPERS**

**3. Apply Machine Learning Algorithms on Big Data**

Your aggregated data tells a complex story. To extract the insights it holds, you need an accurate predictive model that can handle interactions between variables and nonlinearities. You can quickly select and identify the right features for a model and then iterate through additional models to identify the best predictive algorithm.

MATLAB speeds up this process with specialized toolboxes, prebuilt functions, and a full set of statistics and machine learning functionality. It also offers advanced methods such as nonlinear optimization, system identification, and thousands of prebuilt algorithms for financial modeling, portfolio optimization, and risk management (Figure 4).

*"Before we had MATLAB, we would not have been able to produce the clustering model within a reasonable time. We simply would not have done it. MATLAB has opened up new horizons for us."*

**— Pierre-Yves Boillat, Banque Cantonale Vaudoise**

**4. Deploy Machine Learning Models**

MATLAB is an integrated system that allows you to deploy your machine learning models in your production systems as well as integrate with data providers and existing IT infrastructure. Additionally, MATLAB offers online and real-time deployment that integrates into enterprise systems, clusters, and clouds.

*"MATLAB, MATLAB Production Server, and MathWorks Training Services enabled people on our risk team with conditional programming experience in C++ or Java® to efficiently develop a core library for financial analysis and then deploy it as a web application, making it available to production systems in our enterprise environment."*

**— Marcus Veltum, Helaba Invest**

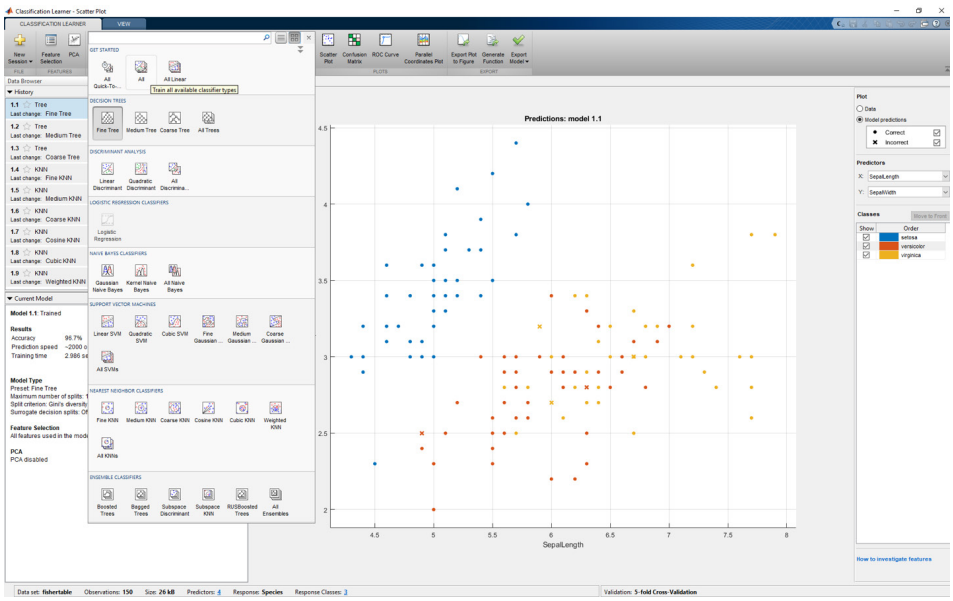



*Figure 4. The Classification Learner app, which lets you interactively train, validate, and tune classification models.*

## Conclusion

Machine learning and big data enable investment managers to make informed decisions based on data-driven insights and predictions not previously available through traditional approaches. New applications and investment insights can be delivered to end customers faster than before.

MATLAB provides an interactive environment and prebuilt functions and libraries to enable quants to become data scientists and develop custom machine learning models. Through flexible deployment options, you can integrate production-ready models quickly into existing IT infrastructure, saving time and eliminating error-prone translations to different programming environments.

## Learn More

*Machine Learning for Algorithmic Trading* (32:55) - Video

*Machine Learning Made Easy* (34:34) - Video

*Forecasting Bitcoin Volatility Using the Regression Learner App* (5:56) - Video

*MATLAB for Quantitative Finance and Risk Management* - Free Product Trial

© 2019 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See mathworks.com/trademarks for a list of additional trademarks.
Other product or brand names may be trademarks or registered trademarks of their respective holders.