Profits and Ethics in Quantitative Trading

An Analysis of the Ethicality and Methods of Leveraging Machine Learning to Predict Future Stock Prices

---

Ryan Mulcahey
5/13/2022 | Washington University in St. Louis | Technical Writing

Abstract

Quantitative researchers and developers interweave financial market data and economic theory to build predictive models. These models are then trained and tested. Only the highest performing models are then used by quantitative traders to reap enormous profits in financial markets. However, unethical means are used in creating and training these models on Wall Street to give traders an edge. I explored various machine learning applications to quantitative trading and various methods of producing price prediction models. I then compared these applications and methods using two criteria: inherent ethical concerns and return on investment. I weighted return on investment more heavily than the inherent ethical concerns as it is generally considered the defining metric of a model, however I certainly addressed the ethical concerns of the methods and models and took them into account in my comparisons. All models explored had strong return on investments, explaining the appeal of the various machine learning methods analyzed. While only the analysis of sentiment analysis based models revealed ethical concerns, analysis of all models and methods displayed the potential for violations of data privacy, based on the choice and use of data. Exploring the intersection of ethics and algorithmic trading, and computing and data science in general, outlined an urgent need for constant changes to data protection and securities regulation as the capabilities of technology grow at an increasing rate. Quantitative trading can be performed unethically given outpaced securities regulation and the focus on profits over ethical data collection and use.

Table of Contents

Introduction

Quantitative trading is an extremely lucrative career path in the financial industry. Quantitative traders, known as quants, are typically recruited from top universities, with math, physics, and computer science degrees. Quantitative traders use their understanding of financial markets and economic theory to pick and choose various predictive models to use in making high speed trades of securities. Available data and chosen algorithms formulate traders' models and strategies. They work in tandem with quantitative researchers and developers who theorize and develop these models [8]. The algorithms produced by these teams and the strategies they use often result in ROI's that outperform the average market, resulting in this lucrativeness of this field. However, profits can come at the cost of ethicality. See Fig. 1 for a visual comparison of how recruiters value the awareness and knowledge of MBA graduates on social responsibility and other issues. One must question the ethical nature of the data collection and usage used in their models, as unethical means are often used on Wall Street to give traders an edge.

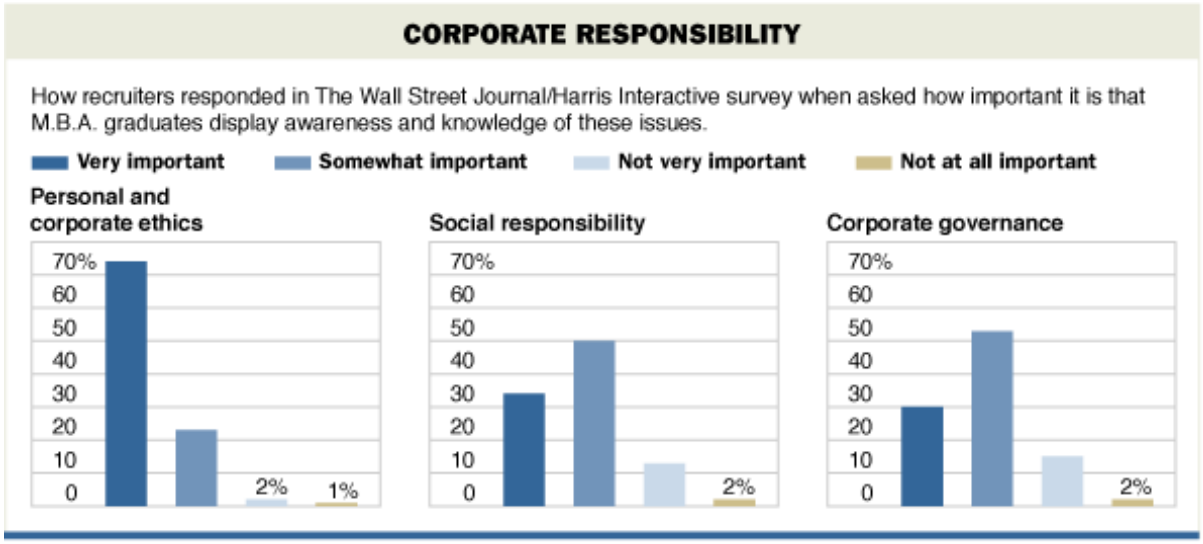

Fig. 1. Recruiters' ratings of the importance of MBA graduates on social responsibility, personal and corporate ethics, and corporate governance. Under 35% of recruiters find a candidate having awareness and knowledge of social responsibility as very important [17]. The implication being that the corporations they represent also do not find social responsibility as a very important issue.

Justification for Research

Those interested in quantitative trading must delve deep into the unethical side of the field, as using unethical means to create a predictive model is wrong. From an ethical point of view, we should not always use all the available data to build our models. See Fig. 2 for a visual

of an ethical data life cycle, with each step of the life cycle broken down into substeps. Some data sets and models take race, socioeconomic status, and other demographic data into account, which could lead to improper conclusions about the data.

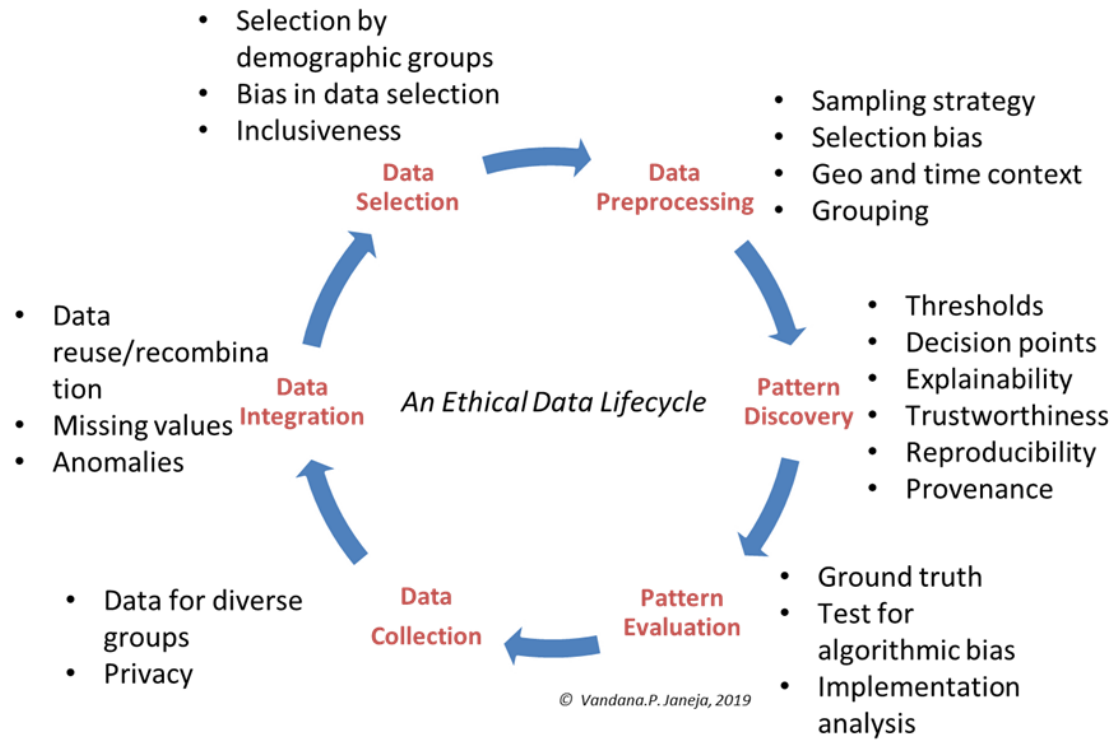

Fig. 2. An ethical life cycle of data selection, cleaning, and use. Ethicality is maintained in all stages of the life cycle from data selection to data integration [13]. Sampling strategy in the data preprocessing stage and privacy in the data collection stage are notable for maintaining an ethical data life cycle.

Privacy concerns regarding data collection must also be taken into account. All data scientists have an inherent obligation to respect the privacy of those they are collecting data on. Additionally, given the high rate of new developments in quantitative trading and machine learning, the field has outpaced securities regulations. See Fig. 3 for a visual of an ethical data life cycle, with each step of the life cycle broken down into substeps. These ethical concerns must be addressed.

Fig. 3. Regulatory solutions to various challenges related to technology. Changes to regulations can match the problems created by evolving technology [15]. Regulations must adapt to evolving technology and should be proactive, so as to not fall behind the rapid growth of technology.

Quantitative Model Life Cycle

The process of developing a price prediction model almost always follows the same process.

1. Identify a problem
2. Data problem
3. Collect data

4. Clean and format data
5. Explore and understand data
6. Use data to create solution
7. Communicate results
8. Repeat if necessary

In a nutshell, this process begins with the theory behind the model. See Fig. 4 for a visual of the basic components of algorithmic trading. Next is the data collection process. And finally, the model is created, trained and tested. While this life cycle is usually consistent for all quantitative models, the methods used by quantitative researchers and developers at each step and the evaluation criteria for each model vary [5]. This life cycle follows the classical model of collecting data based on a theory, versus collecting data to then create a theory. The latter of which falls under data mining.

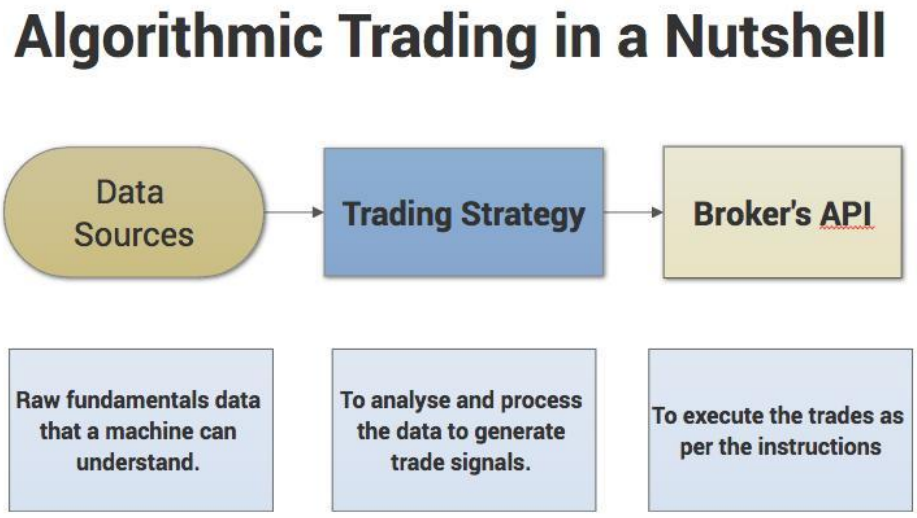

Fig. 4. The basic components of algorithmic trading. Understandable data sources are needed to formulate a trading strategy, which will then be automated for a trader to execute on real financial markets [14]. Financial market and economic theory is needed, then data is needed to develop a trading strategy based on the theory.

Models vary for two reasons. The methods used to develop them and their data sources. Ethical pitfalls can be mediated or introduced in both the choice of data sources and their use in model development. However, traders must remember that it is possible to profit in an ethical manner. See Fig. 5 for a visual of the intersection of economic value and ethical value. While the development of models and the choices made in their development vary, the statistical tools used in their development are limited to a smaller number of platforms, most commonly R, Python and MATLAB [1]. Within the chosen platform, a quantitative developer begins by loading in the

chosen data for a model. A developer must then have a strong understanding of the data in order to use this. For example, to obtain this understanding using R, a developer would need to install various statistical software packages to parse the data and turn it into an understandable form for both the developer and R to understand [3]. Once this data processing is completed, a developer could see the summary statistics and other readable information about the dataset to build a strong understanding of the data they are using.

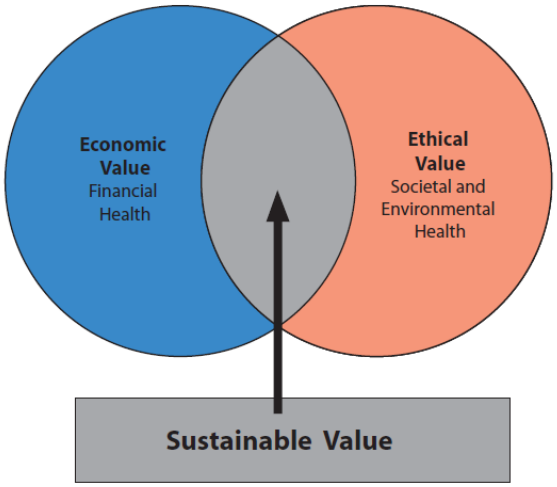

Fig. 5. A Venn diagram of economic value and ethical value. The intersection of economic value or profit and ethical value is sustainable value [12]. Quantitative traders should strive for the intersection of ethicality and profits.

Analysis of Machine Learning Methods

There are two overarching strategies to develop a quantitative model. The first is the classical model, which follows the classical quantitative model life cycle. This involves finding data to find a model based on financial market and economic theory. The second is the data mining approach, where data is identified first and connections and relationships within the data are used to formulate the model. While the applications of data mining in the financial industry are compelling, this analysis will solely look at the models that fall under the first category, as one can not accurately compare the quality of models whose underlying frameworks are polar opposites.

Momentum Strategy Derived from Big Data Analysis:

The theory behind momentum driven trading strategies differs greatly from other strategies. Put simply, momentum strategies capitalize on market volatility. This phenomenon is

the constant change in stock prices and provides traders with the opportunity to make or lose money. The idea of buying low and selling high seems fairly straightforward. However, big data gives traders the ability to better time their trades [2]. The past decades have seen widespread adoption of big data analysis in quantitative trading, as access to financial data has become widely available. This increase in data collection and availability is associated with traders being able to better calibrate their models. Using big analysis, Gao et al. created four quantitative models based on momentum strategies [2]. In their study they found an apparent tradeoff between increasing the risk of their portfolio with decreasing the diversification of their portfolio. The return on investments of their four models depended on the balance between these two competing factors. The life cycle of these models lacked any ethical breaches of data use.

Sentiment Analysis:

Sentiment analysis is the process of parsing through user generated data to identify a sentiment score. A sentiment score is typically based on how many positive words or phrases appear in a review or article, in relation to the number of negative words or phrases. Applying sentiment analysis to a trading strategy is a theoretically sound decision, as the emotions of investors play a role in determining the prices of securities [6]. However, applying sentiment analysis to a trading strategy is difficult at the technical level. Data from a representative sample of investors must be collected.

Li et al. turned to social media as their data source to build a media-aware quantitative trading strategy. Analyzing tweets from Twitter and economic data and news articles from other sources, they concocted overall emotion and public mood scores. These scores were then applied to various supervised learning models, which gave these researchers predicted future stock prices. These supervised learning models included basic linear regression models, as well as more complicated models, such as k-nearest neighbor models. The quantitative models created from this process correctly predicted the movement of future stock prices 55.08% of the time [6]. The framework of using news articles to predict future stock prices outlines the immense power that news networks possess. Investors must question the ethicality of this power being in the hands of so few. The success of these quantitative models based on sentiment analysis reveal that the opinion of new networks on the state of financial markets directly impacts investors' emotions, and thus the prices of securities.

Online Learning:

Machine learning models generally use a predefined data set, which is split into training and testing data. The training data is used to train the model, whereas the testing data is used to test the model generated by the training data to determine the model’s accuracy or predictive power. Online learning models follow a different course of action in their data selection. An online learning model continuously reassesses itself as new data becomes available, and adjusts

itself accordingly, rather than just once during the training phase, as traditionally done following the generic offline method.

Historically, asset portfolios were selected to optimize the expected return over a specified period of time, at an accepted level of risk. Recently, investors and portfolio managers have found another option, online portfolio selection. This newfound portfolio selection process allows investors to optimize the expected return continuously over multiple periods of time, at their accepted level of risk. This second approach to portfolio selection, using online learning, is the optimal long-term selection strategy [7]. While online learning does not introduce any ethical concerns that are not found in its alternative, offline learning, one must acknowledge the violations of data privacy found in the data life cycle of both online and offline learning models.

Reinforcement Learning:

Reinforcement learning is a subset of machine learning, with applications in statistics, game theory, and quantitative trading. Reinforcement learning algorithms train an artificial intelligence or model to make accurate decisions by adjusting the decision maker, either the artificial intelligence or model, based on the accuracy of its decisions. Positive actions or decisions are rewarded, whereas negative performance results in a negative consequence. In practice, training an algorithm through reinforcement learning will result in a selection model that can give an expected return for securities selections made by quantitative traders.

Zhang et al. devised a selection model trained on data from the Chinese Commodity Futures Market that outperformed baseline algorithms both in terms of profit and time cost [11]. In this context, time cost is the amount of time that it takes an algorithm to reach a decision after receiving data. The reinforcement learning based selection algorithm outperformed the baseline algorithms by a factor of ten in relation to this metric [11]. Zhang et al.'s application of reinforcement learning to quantitative trading displays how it is not any particular quantitative method that violates ethics in quantitative trading. Ethical violations come about when actors either unethically or ignorantly choose to violate data privacy or securities regulations. Neither of these violations occurred in Zhang et al.'s research, as they simply used generic data on a futures market.

Cluster Analysis:

Reinforcement and online learning are methods that can be applied to any machine learning model. Machine learning models can also be either supervised or unsupervised. However, at their core, machine learning models fall under three distinct categories: classification, regression, and clustering. Classification models use input data to determine which class a specific data point falls under. On the other hand, regression models predict a value for the target variable of a data point based on input data. These two categories of machine learning models are supervised, meaning input data is used to predict the value of the dependent variable

for each data point. Meanwhile, clustering models are unsupervised, meaning that it uses data without any specific dependent feature.

Clustering models attempt to group data points into a specific number of groups, or clusters, based on some metric of similarity. Clustering has apparent applications to quantitative trading, as it allows traders to visualize whether a specific stock shares similarities with poorly-performing or well-performing stocks. In practice, a K-means clustering model developed by Wu outperformed the S&P 500 index in both bear and bull markets. Wu divided American stocks into multiple clusters and then constructed a stock portfolio, consisting of the most-centered stocks in the best-performing clusters [9]. Once again, the ethicality of this machine learning application to quantitative trading is dependent on the way data is used and the choice of data. As Wu simply used basic financial data on American stocks, Wu's application of cluster analysis to quantitative trading has no ethical concerns.

Graph-Based Stock Correlation and Prediction:

Stock prices are influenced by a seemingly infinite number of variables. Traders have always worked to account for as many of these variables as possible in their trading models. Yin et al. created a Graph Attention Long Short-Term Memory machine learning model to attempt to correct for omitted variables in traditional models [10]. In equilibrium theory, investors work under the assumption that financial markets are perfect markets. However, financial markets are imperfect, as investors leverage imperfect pricing to profit. Despite imperfect pricing, investors can use previous stock prices to predict future stock prices.

This concept is the basis for the machine learning model developed by Yin et al. After computing a correlation matrix to see the correlations between individual stocks over time, they developed their machine learning model, which uses these correlations to predict the future prices of stocks. In 13 weeks of testing on the Chinese A-share market, their model garnered a ROI of 44.71% [10]. As seen with other machine learning methods, it is the choice and use of data which determines the ethicality of a method. Similar to previously discussed models, Yin et al.'s model lacked any ethical concerns, as it was only trained with basic financial data of the Chinese A-share market.

Conclusions: Securities Regulations and Data Protection Laws

Lawmakers and regulars must update securities regulations and data protection laws to account for the potential ethical implications of the applications of different machine learning methods. While it is not necessarily machine learning methods that lead to violations of data privacy, the choice and use of data in the training of specific models using these methods can be unethical. Furthermore, the staggering return rate of the sentiment analysis based model should lead the average investor to question the power media corporations hold over the pricing of

securities. Widespread adoption of profitable algorithmic trading has given traders the means to profit off of the violation of data privacy [4]. While quantitative traders and developers may not always intend to violate data privacy, temptations for profit and the mere means to violate data privacy warrant a closer look at today's securities regulations and data protection laws.

Regulations and laws must be updated to account for the advancing bounds of technological capabilities. As automated high-frequency trading algorithms continue to execute an increasing percentage of securities trades, false new reports or over-hyped information have an increasing influence on the state of financial markets [16]. The power held by social media platforms and media corporations, on which this falsified or overhyped news is spread, must be checked by updated securities regulations. Technological evolution and new methods of trading must be met by equally evolving laws and regulations.

References

[1] Chan, Ernest. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business*, 20 May 2021, https://books.google.com/books?hl=en&lr=&id=j70yEAAAQBAJ&oi=fnd&pg=PR1&dq=%22quantitative+trading%22&ots=wsOl7VXITP&sig=ebhqTPTNSI8oNsKRjEk3d2qelY8#v=onepage&q=%22quantitative%20trading%22&f=false.

[2] Gao, X., et al. *Big Data Analysis with Momentum Strategy on Data-driven Trading*, 30 Sept. 2021, https://www.scopus.com/record/display.uri?eid=2-s2.0-85124120549&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=22&citeCnt=0&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[3] Georgakopoulos, H. *Quantitative Trading with R: Understanding Mathematical and Computational Tools from a Quant's Perspective*, 2 Feb. 2015, https://www.scopus.com/record/display.uri?eid=2-s2.0-84973460171&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=104&citeCnt=2&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[4] Kunz, K., Martin, J. *Into the Breech: The Increasing Gap between Algorithmic Trading and Securities Regulation*, Feb. 2013, https://www.scopus.com/record/display.uri?eid=2-s2.0-84886383038&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=107&citeCnt=3&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[5] Lauprete, G. *Alpha Genesis - The Life-Cycle of a Quantitative Model of Financial Price Prediction*, 4 Sept. 2015, https://www.scopus.com/record/display.uri?eid=2-s2.0-84984861416&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=103&citeCnt=0&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[6] Li, Q., et al. *Media-aware quantitative trading based on public Web information*, May 2014, https://www.scopus.com/record/display.uri?eid=2-s2.0-84897550633&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=117&citeCnt=51&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[7] Li, Y., et al. *When quantitative trading meets machine learning: A pilot survey*, 9 Aug. 2016, https://www.scopus.com/record/display.uri?eid=2-s2.0-84986601913&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=102&citeCnt=5&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[8] Ta, Van-Dai, et al. *Prediction and Portfolio Optimization in Quantitative Trading Using Machine Learning Techniques*, Dec. 2018, https://dl.acm.org/doi/pdf/10.1145/3287921.3287963.

[9] Wu, S. *Application of Cluster Analysis in Stock Selection in United States Stock Market*, 1 Oct. 2020, https://www.scopus.com/record/display.uri?eid=2-s2.0-85096121646&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=38&citeCnt=0&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[10] Yin, T., et al. *Graph-based stock correlation and prediction for high-frequency trading systems*, Feb. 2022, https://www.scopus.com/record/display.uri?eid=2-s2.0-85113573353&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=4&citeCnt=3&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[11] Zhang, W., et al. *Auto uning of price prediction models for high-frequency trading via reinforcement learning*, May 2022, https://www.scopus.com/record/display.uri?eid=2-s2.0-85123364176&origin=resultslist&sort=plf-f&src=s&st1=%22quantitative+trading%22&nlo=&nlr=&nls=&sid=3dc3d65e48d0d46d58509a677d8cb18a&sot=b&sdt=b&sl=37&s=TITLE-ABS-KEY%28%22quantitative+trading%22%29&relpos=0&citeCnt=0&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.

[12] “Building Sustainable Value through Fiscal and Social Responsibility.” *Ivey Business Journal*, 12 Dec. 2011, https://iveybusinessjournal.com/publication/building-sustainable-value-through-fiscal-and-social-responsibility/.

[13] “Do No Harm: An Ethical Data Life Cycle.” *AAAS Science & Technology Policy Fellowships*, 4 Apr. 2019, https://www.aaaspolicyfellowships.org/blog/do-no-harm-ethical-data-life-cycle.

[14] “Fundamental Analysis with Algorithmic Trading.” Quantitative Finance & Algo Trading Blog by QuantInsti, Quantitative Finance & Algo Trading Blog by QuantInsti, 27 Jan. 2021, https://blog.quantinsti.com/fundamental-analysis-performed-algorithmic-trading/.

[15] “National Security and Technology Regulation.” *Deloitte Insights*, 12 July 2019, https://www2.deloitte.com/us/en/insights/industry/public-sector/national-security-technology-regulation.html.

[16] “Regulators, Private Investors Outpaced by Algorithmic Stock Trading.” *Vanderbilt University*, 29 Apr. 2014, https://news.vanderbilt.edu/2014/04/29/humans-outpaced-by-algorithms/.

[17] “Right and Wrong.” *The Wall Street Journal*, 17 Sept. 2003, https://www.wsj.com/articles/SB106365505376228100.