

# Putting the Power of AI to Work in Investing

**Minkyu Kim, Ph.D.**

Managing Director, Head of Data and Analytics  
Systematic Equity

**Alexander Rudin, Ph.D.**

Managing Director  
Head of Multi-Asset and Fixed Income Research

**Gaurish Agrawal**

Senior Quantitative Research Analyst  
Systematic Equity

ChatGPT, the artificial intelligence chatbot from OpenAI, has become a global phenomenon, attracting more than 100 million active users in just two months after its launch — a record for the fastest-growing consumer application in history.<sup>1</sup> The technological models powering ChatGPT, advanced machine learning algorithms, have made rip-roaring waves in a wide range of fields, from protein structure prediction in biology to creator tools in social media. In this piece, we consider how AI and machine learning<sup>2</sup> could apply to investment decision-making. It turns out that integrating AI into the financial system requires thoughtfulness, as finance does not neatly meet the criteria of categories that are compatible with AI.

**A glossary of helpful terms can be found at the bottom of page 8.**

---

**Compatibility With  
Machine Learning**

---

Machine learning is a paradigm shift from traditional science. Unlike the traditional scientific method that involves forming a hypothesis and testing it to draw conclusions, machine learning uses statistical techniques to uncover patterns in data to find answers; no hypothesis or explicit understanding of the system is required. As a result, machine learning can find an open door to questions that are too multifaceted for humans to fully understand, or that are too time-consuming for consumers — or companies — to figure out.

---

Often, discerning the patterns from data — a key strength of machine learning — is enough to solve many real-life problems. For example, in the fields of natural language processing and image processing, machine learning has been so highly effective that it has fundamentally transformed the way things are done. However, since AI is domain-agnostic at the algorithm level, we can surmise that certain shared characteristics of those systems are conducive to finding answers by pattern detection.

Those characteristics are:

- **Stability in the System** Language is an example of a stable system, in that the rules and structure of the language typically remain consistent while, and after, the algorithm learns them.
- **Abundant Data** A large amount of clean data is available for the algorithm to learn the complex systems during the “training” stage.<sup>3</sup>
- **Intuitive Outcomes** The outputs from machine learning make intuitive sense, so it is less necessary for the user to understand *why* the model made a certain determination.

For systems with the above traits, we conclude that if a machine-learning-based model is rigorously trained and validated in a testing environment, it will also perform well with yet-to-be-seen live data.

---

## Relevance to the Financial System

---

We considered whether the above traits apply to the financial system, and tested the crucial assumption that machine learning models, if well-trained and validated, will perform similarly when making investment decisions in an out-of-sample, or “real world,” scenario. In our view, several aspects of the financial system make it less apt for the integration of AI:

- **Financial Markets Lack Stability** Financial markets change frequently, with outcomes driven by dynamic interactions among market participants. Variables in the financial system and the relationships between them often experience structural shifts.
- **Financial Information is Finite** As opposed to other systems such as language, financial input is more limited in quantity. The amount of data available for modeling depends on the frequency and horizon of interest — the shorter the horizon, the easier to collect a large amount of data — but we are fundamentally bound by what has happened in the market already, as market data cannot be newly manufactured. Also, for any market data indicative of future asset returns, the signal-to-noise ratio is typically very low — most of the predictive information available in efficient markets is “noise” because useful information has largely been arbitrated away.
- **Financial Model Outputs are Complex and Require the Passage of Time to Evaluate** It is rarely straightforward to determine whether investment decisions are prescient or detrimental until time passes, making model outputs difficult to evaluate. In addition, prudent investors will want to understand why a model made a certain decision before they act upon it. Furthermore, fiduciaries may seek to explain their investment decisions to clients, or they may face regulatory requirements to do so.

At bottom, we do not believe that machine learning models, unless carefully designed with investment expertise and executed within a rigorous research discipline, will predict favorable investment decisions with the same accuracy in an out-of-sample scenario. Due to the above challenges, the risk is high that machine learning models may “overfit” investment data, meaning that the model is set to fit too closely to the training data and therefore does not generalize to unseen data. In our view, an entirely data-driven approach — as performed by typical machine learning applications — is unlikely to succeed when tackling investment problems.

---

## Where Machine Learning Can Provide Useful Financial Insights

To derive tangible benefits from AI in client portfolios, which is, after all, the primary goal of any financial model, the keys are: (1) to understand the conditions required for machine learning to be successful, and (2) to apply the model's outputs in the right way. For a prudent investor, the calls to integrate AI could prompt a journey to find the right level of trade-off between the traditional modeling culture, where major decisions are made by human analysts, and the data-driven culture of machine learning.

For the financial system, we considered steps that may allow investors to overcome the three challenges (system stability, data availability and quality, and ease of understanding outputs) listed above.

- **Focus on a Stable, Persistent Phenomenon in Finance** Rather than aiming to model the full dynamic and complex financial system at once, we recommend a more measured approach that focuses on relatively stable and time-invariant relationships between input and output variables. Of course, the exact relationship doesn't need to be formulated as in traditional modeling, but the success rate will be higher for a pattern that can reasonably be assumed to persist due to information asymmetry, behavioral biases, or market frictions.
- **Adjust the Model Per Data Quantity and Quality** When the amount of data is limited, the complexity of the model needs to be adjusted accordingly. For example, decision trees, one type of machine learning algorithm, will need a lot less data than neural networks, a newer and more advanced type of algorithm. In addition, the quality of data can often be significantly enhanced by "feature engineering," a process by which input data is carefully treated and converted using domain expertise.
- **Apply Risk Controls** To overcome the challenge of limited explainability, an appropriate risk control measure should be in place. For instance, if machine learning is used as a subcomponent within a structured model with clear boundaries and limited risk budget, its non-transparent nature will be easier to deal with. Or, if the output of a machine learning model is subject to validation by human analysts, explaining the reason behind the output may not be crucial.
- **Weigh Against the Incumbent Model** If there's already a structured model with an established optimality criterion, switching to machine learning will typically require careful balancing of improved performance and less transparency. When the current best approach is mere heuristics or there's simply no alternative, machine-learning-based models are often a reasonable choice.

## Case Studies

### Harvesting Information from Unstructured Data

Below, we review a selection of cases in which the Systematic Equity; Investment Solutions Group (ISG); and Fixed Income, Cash, and Currency (FICC) teams have used machine learning to improve investment outcomes in client portfolios.

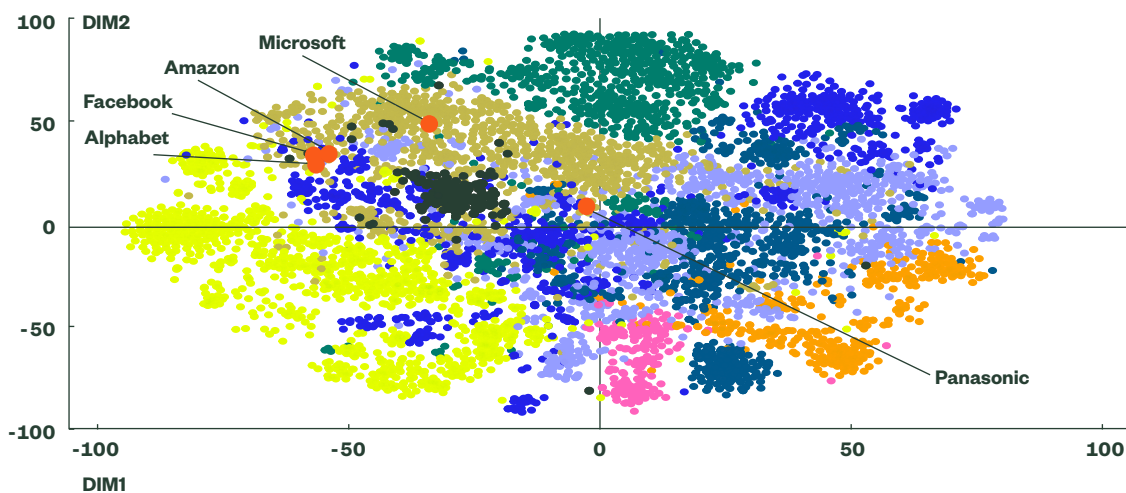
The availability of unstructured data has increased in recent decades, with over 80% of data available on the internet being unstructured.<sup>4</sup> In finance, unstructured textual data such as earnings call transcripts or regulatory filings has been a relatively untapped source of information, but it can offer valuable insights into a company's operations and business prospects.

The Systematic Equity team employs a suite of factors based on natural language processing (NLP). These factors are designed to capture return-influencing features and insights from unstructured textual data. For earnings call transcripts, for instance, they apply a variety of techniques from traditional linguistic processing to advanced machine learning to analyze a wide range of measures such as the tone and complexity of the language used, the subtle sentiment behind executives' statements, and the management's behavior during the call (see *Extracting Complex Investment Insights from Earnings Calls*). These combined measures proved to enhance the Systematic Equity team's ability to assess the overall sentiment and quality of announced results beyond the immediate financial results released in each reporting cycle. Other useful textual data the Systematic Equity team processes includes regulatory filings, patents, job postings, etc.

One of the advanced machine learning techniques used by the Systematic Equity team is a neural-network-based "embedding" that converts the textual data into lower-dimensional numerical vectors, preserving semantic and syntactic information. Once the texts are converted, the companies that discussed similar topics will appear clustered together in the resulting numerical vector space (Figure 1), which can serve as a useful byproduct to identify novel peer groups beyond the traditional sector/industry classifications.

Figure 1  
**NLP Can Help Group Companies Discussing Similar Topics on Earnings Calls**

- Communication Services
- Communication Staples
- Financials
- Industrials
- Materials
- Consumer Discretionary
- Energy
- Health Care
- IT
- Utilities



Source: State Street Global Advisors, Company Reports as of May 31, 2023. T-SNE applied to document embeddings.

---

In the above figure, each point represents an earnings call transcript that was published in May 2018. For each transcript, 300-dimensional embedding was generated using a neural-network-based embedding technique. Embeddings were then projected onto the two-dimensional space for ease of visualization. While most of the companies belonging to the same sector are clustered together, there are also clusters consisting of companies from multiple sectors. One example is Amazon (Consumer Discretionary), placed much closer to Alphabet, Facebook (Communication Services) and Microsoft (Information Technology) than its sector peer Panasonic.

A similar machine learning technique can also be used to identify companies relevant to certain “themes,” such as particular technologies or macroeconomic trends, by processing the filings, earnings call transcripts, and various disclosures from the companies.

Overall, NLP is an ideal way to apply machine learning in finance, as one can train the model using a large corpus of textual data without relying on noisy market data or even borrow pre-trained language models directly from the mature field of NLP machine learning research. And the core language model will remain stable even when its outcomes may need to be used differently in response to changing market conditions.

---

## Modeling Beyond a Static Linear Relationship

In tactical asset allocation (TAA) (see *Why Institutional Investors Should Consider Tactical Asset Allocation*), the relationship between asset prices and underlying factors is often complicated and far from a simple linear dependency. ISG has been a pioneer in what we refer to as regime-driven tactical investing (see *State Street Global Advisors Market Regime Indicator: Q422*), which is based on assessing a market environment’s “regime,” or level of risk aversion. The ISG team combines the regime indicator models with machine learning techniques such as hidden Markov models (HMMs) to evaluate the relative attractiveness of various assets and forecast total returns.

The ISG team employed HMMs first in 2004 as part of their process to forecast commodities prices, and later in 2010 to tactically allocate between emerging and developed equity markets. HMM remains a technique the ISG team employs, when appropriate. In 2022, they developed a new model for predicting credit spreads<sup>5</sup> that employs an HMM. The ISG team plans to deploy it as part of their TAA strategy in 2023.

The ISG team has also been using dynamic linear modeling for situations in which the time variation of asset price to factor relationship is material, but gradual. Such machine learning techniques as the Kalman Filter have also been utilized by the ISG team since 2014 in application to fixed income sector rotation problems and in their hedge fund replication strategies. Finally, in 2022, the ISG team investigated if a decision-tree-based approach is superior to a simple linear model when using macro data changes as factors for tactical equity positioning; early results look promising.

The Systematic Equity team has developed an XGBoost-based method to forecast company fundamentals such as earnings and cash flows, crucial inputs for equity investors in determining the expected return of a company stock. The machine learning method, which allows the modeling of complex relationships among the trailing firm fundamentals, fast-moving market variables, analyst estimates, and macro indicators, proved its clear advantage in terms of accuracy over the plain sell-side analyst estimates across multiple regions. While the method has not been implemented in live portfolios, as overlapping information was already present in other parts of the Systematic Equity team’s model, machine-learned company fundamentals can serve as useful inputs in other contexts, providing in-house estimates with better coverage and fewer biases than conventional analyst estimates.

---

## Reducing Estimation Noise

The relatively long time horizon of tactical investing, combined with the relatively small number of available degrees of freedom, makes reducing estimation noise a key focus of our efforts. The ISG and FICC teams made early attempts in employing such machine learning techniques as regularized regressions (such as LASSO, probit, or others) more than a decade ago. More recently, those teams introduced additional techniques, some of which they developed internally. In 2019, the ISG team incorporated random correlation matrix cleaning into our portfolio construction process. In 2020, they suggested an adaptive enhancement to a classical risk parity strategy<sup>6</sup> that can be viewed as a machine learning approach. In 2021, they conceptualized on how to incorporate fuzzy mathematics — a well-known AI technique — into portfolio optimization as an antidote to the estimation noise.<sup>7</sup> Currently (March 2023), the ISG team is working on a risk model that uses PAM clustering as a way to intelligently reduce the complexity of the emerging market bond universe.

---

## Conclusion

Much ink has been spilled about AI and its applications across multiple categories. For financial systems, AI can be useful in offering insights from sources that are infrequently tapped, modeling asset prices and factors with nonlinear relationships, or reducing estimation noise, among other uses. What's important, however, is understanding the contexts in which machine learning outcomes are the most helpful and reliable. In our view, machine learning is best when combined with the knowledge of strong investment teams that can control the structures, contexts, and training and determine the questions that AI will address. When used thoughtfully, AI can be a powerful investment tool to unlock complex, previously untapped information from a wide range of variables.

---

## Endnotes

- 1 L. Walmsley, C. Kuntarich, K. Madhukar, R. Freeman and E. Vaish, "US Internet: ChatGPT officially crosses the 100M MAU mark in January," UBS Global Research and Evidence Lab Report, Feb. 2023.
- 2 Machine learning is the science of programming computer systems to learn from data for the purpose of prediction or decision-making, whereas AI refers to systems, employing machine learning and/or other techniques, that are able to perform tasks that ordinarily require human intelligence. Throughout this article, though, AI and machine learning will be used interchangeably.
- 3 Even a specialized algorithm designed to understand specific expressions used in a narrow field (e.g., equity investing), which limits the amount of available data, can still use the full panoply of textual data to learn the language itself (a step called "pre-training"), and then only a fractional amount of domain-specific text is required later for "fine-tuning."
- 4 The International Data Corporation, 2020.
- 5 Pravesh Kumar, Rahul Sathyajit, and Alexander Rudin, "Regime Switching With Gradual Transition — A New Model For Credit Spreads and Its Applications to Tactical Asset Allocation," ISG working paper.
- 6 Alexander Rudin, Vikas Mor, and Daniel Farley, "Adaptive Optimal Risk Budgeting," *The Journal of Portfolio Management*, Multi-Asset Special Issue 2020, 46 (6) 147–158.
- 7 Alexander Rudin and Daniel Farley, "Fuzzy Factors and Asset Allocation," *The Journal of Portfolio Management*, Multi-Asset Special Issue 2021, 47 (4) 110–122.

## About State Street Global Advisors

For four decades, State Street Global Advisors has served the world's governments, institutions, and financial advisors. With a rigorous, risk-aware approach built on research, analysis, and market-tested experience, we build from a breadth of index and active strategies to create cost-effective solutions. As pioneers in index and ETF investing, we are always inventing new ways to invest. As a result, we have become the world's fourth-largest asset manager\* with US \$4.37 trillion<sup>†</sup> under our care.

\* Pensions & Investments Research Center, as of December 31, 2023.

<sup>†</sup> This figure is presented as of June 30, 2024 and includes ETF AUM of \$1,393.92 billion USD of which approximately \$69.35 billion USD is in gold assets with respect to SPDR products for which State Street Global Advisors Funds Distributors, LLC (SSGA FD) acts solely as the marketing agent. SSGA FD and State Street Global Advisors are affiliated. Please note all AUM is unaudited.

### ssga.com

**Marketing communication.**  
For investment professional use only.

#### Glossary

**Machine Learning** The science of programming computer systems to learn from data for the purpose of prediction or decision-making.

**Artificial Intelligence** Systems, employing machine learning and/or other techniques, that are able to perform tasks that ordinarily require human intelligence.

**Hidden Markov Model** A statistical model that uses observations to infer the probability of a hidden state, assuming that the hidden state is a Markov process with unknown parameters that need to be estimated.

**Kalman Filter** An algorithm that uses a series of measurements over time to estimate the true value of a variable, while also taking into account measurement noise and uncertainty.

**LASSO (Least Absolute Shrinkage and Selection Operator)** A linear regression model that performs variable selection and regularization by adding a penalty term to the sum of squared errors in the regression equation.

**Probit** A type of regression analysis that models the relationship between a binary response variable and one or more predictor variables using the cumulative distribution function of a standard normal distribution.

**Degrees of Freedom** The number of independent observations in a sample that are

free to vary after certain constraints, such as the sample mean, have been calculated.

**Estimation Noise** The variability or uncertainty in the measurements or observations used to estimate a parameter or model, which can result in errors or inaccuracies in the estimation process.

**Random Correlation Matrix** A square matrix in which each entry represents the correlation between two variables and is generated randomly, subject to the constraints that the matrix is positive definite and has a diagonal of ones.

**Risk Parity Strategy** An investment portfolio optimization technique that seeks to allocate portfolio weights in such a way that each asset class contributes equally to the portfolio's overall risk, rather than equally to its return as in a traditional portfolio.

**Fuzzy Mathematics** A branch of mathematics that deals with uncertainty and imprecision by allowing values to be expressed in degrees of truth rather than absolute values, enabling more flexible and nuanced analysis and decision-making.

**PAM (Partitioning Around Medoids) Clustering** A clustering algorithm that aims to group similar data points together by iteratively replacing the mean (centroid) of a cluster with a more representative medoid, or most centrally located data point, until a stable set of clusters is obtained.

**XGBoost (eXtreme Gradient Boosting)** A machine learning algorithm that uses an ensemble of decision trees, where each tree is trained on the residuals of the previous tree, to iteratively minimize a loss function and improve prediction accuracy.

### State Street Global Advisors Worldwide Entities

#### Important Risk Information

The information provided does not constitute investment advice and it should not be relied on as such. It should not be considered a solicitation to buy or an offer to sell a security. It does not take into account any investor's particular investment objectives, strategies, tax status or investment horizon. You should consult your tax and financial advisor.

This document contains certain statements that may be deemed forward-looking statements. Please note that any such statements are not guarantees of any future performance and actual results or developments may differ materially from those projected. Investing involves risk including the risk of loss of principal.

**The information contained in this communication is not a research recommendation or 'investment research' and is classified as a 'Marketing Communication' in accordance with the Markets in Financial Instruments Directive (2014/65/EU) or applicable Swiss regulation. This means that this marketing communication (a) has not been prepared in accordance with legal requirements designed to promote the independence of investment research (b) is not subject to any prohibition on dealing ahead of the dissemination of investment research.**

The views expressed are the views of the Minkyu Kim, Ph.D., through September 10, 2024, and are subject to change based on market and other conditions.

All information has been obtained from sources believed to be reliable, but its accuracy is not guaranteed. There is no representation or warranty as to the current accuracy, reliability or completeness of, nor liability for, decisions based on such information.

Investing involves risk including the risk of loss of principal.

The trademarks and service marks referenced herein are the property of their respective owners. Third-party data providers make no warranties or representations of any kind relating to the accuracy, completeness or timeliness of the data and have no liability for damages of any kind relating to the use of such data.

Equity securities may fluctuate in value and can decline significantly in response to the activities of individual companies and general market and economic conditions.

Bonds generally present less short-term risk and volatility than stocks, but contain interest rate risk (as interest rates rise, bond prices usually fall); issuer default risk; issuer credit risk; liquidity risk; and inflation risk. These effects are usually pronounced for longer-term securities. Any fixed income security sold or redeemed prior to maturity may be subject to a substantial gain or loss.

© 2024 State Street Corporation.  
All Rights Reserved.  
ID2352801-5729120.5.4.GBLINST 0924  
Exp. Date: 09/30/2025