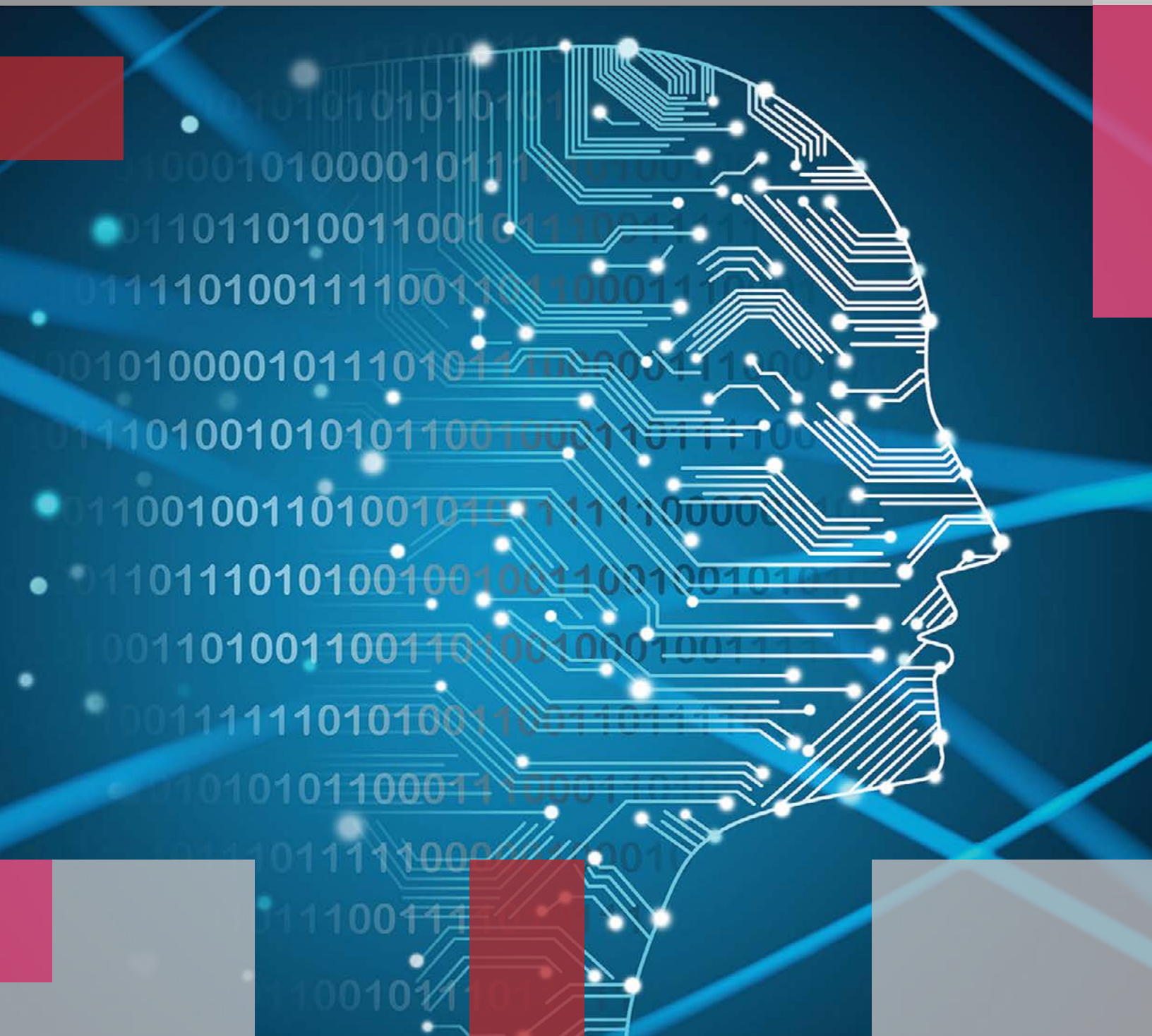
CFA Institute

AI PIONEERS IN INVESTMENT MANAGEMENT

An examination of the trends and use cases of AI and big data technologies in investments



DISCLAIMER

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is distributed with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. Furthermore, the publisher is not providing investment advice or endorsing any methodology. The cases included in the report are not an endorsement of the people, products, or firms mentioned in the cases; investment product returns are unpredictable. If expert assistance is required, the services of a competent professional should be sought.

ISBN: 978-1-942713-78-4

CE Qualified Activity CFA Institute

This publication qualifies for 1.25 CE credits under the guidelines of the CFA Institute Continuing Education Program.

AI PIONEERS IN INVESTMENT MANAGEMENT

CONTENTS

EXECUTIVE SUMMARY

Will robots replace human investment managers? As the investment industry stands on the cusp of arguably its greatest technological transformation, we set out to understand the current state of adoption of artificial intelligence (AI) and big data applications in investment management and to exemplify where and how such technologies can be put to use.

We found that relatively few investment professionals are currently exploiting AI and big data applications in their investment processes. To provide a guidepost for investment firms and individuals seeking to move toward the latest technological frontier, we spoke with a selection of institutions across the globe that are currently using these technologies; these are among the AI pioneers in investment management.

Their use cases, presented in this report, are illuminating. Among other things, they underscore the opportunities but also the limitations of AI and the continued important role of human judgment in investment processes.

We ascribe to the power of the "AI + HI" model: AI techniques can augment human intelligence to enable investment professionals to reach a higher level of performance, freeing them from routine tasks and enabling smarter decision making that leverages the collective intelligence of machines and humans.

Successful investment firms of the future will be those that strategically plan on incorporating AI and big data techniques into their investment processes. Successful investment professionals will be those who can understand and best exploit the opportunities brought about by these new technologies.

The future is here.

HIGHLIGHTS

- We identify three types of AI and big data applications in investment management: (1) using natural language processing (NLP), computer vision, and voice recognition to efficiently process text, image, and audio data; (2) using machine learning (ML), including deep learning, techniques to improve the effectiveness of algorithms used in investment processes; (3) using AI techniques to process big data, including alternative and unstructured data, for investment insights.
- According to a CFA Institute survey, relatively few investment professionals are currently using AI/big data techniques in their investment processes. Most portfolio managers continue to rely on Excel and desktop market data tools; only 10% of portfolio manager respondents have used AI/ML techniques in the past 12 months.
- We identify five major hurdles to successful adoption of AI and big data in investment processes: cost, talent, technology, leadership vision, and time. Investment firms will need to substantially overcome the five hurdles to reach the top of the FinTech pyramid.
- Powerful FinTech will be the result of collaboration between Fin (financial institutions) and Tech (technology companies). Successful firms will be centered on T-shaped teams that combine investment expertise, innovation, and technology application across investment strategies or processes.

Key Takeaways

- The decision to use AI and big data techniques should be benchmarked against the performance of traditional techniques. Firms should determine whether the potential additional alpha capture is worth the additional cost and complexity of applying AI and big data.
- A machine is only as intelligent as the data it learns from. The more comprehensive the training data, the more generalized the machine will process new events, thereby mitigating common pitfalls like overfitting.
- ML techniques are more suited to systematic strategies (including rules-based, quantitative strategies), and unstructured and alternative data are typically used more by discretionary (active) managers.
- Niche, sector-specific data sets are of more relevance to a fundamental analyst or portfolio manager searching for alpha than a systematic manager.
- The effective use of such data sets could provide one of the biggest opportunities for a besieged active management sector.
- AI and big data are no panacea; they cannot solve every investment problem. For example, only a small proportion of big data can generate meaningful signals; reliably extracting signal from noise is difficult.

INTRODUCTION

In this report, we seek to identify high-impact applications of artificial intelligence (AI) and big data in investments and best practice in their implementation by examining specific use cases. For this purpose, we conducted interviews with investment industry practitioners around the world and from different areas of investments, mostly in April and May 2019.

Why Are We Launching This Report?

As noted in previous CFA Institute research initiatives—most recently, *Fintech 2018: The Asia Pacific Edition*[1]—we believe that the early stages of FinTech (e.g., peer-to-peer lending, mobile payment, and robo advice) are more complementary than disruptive to the financial services industry. However, new technologies, such as the ABCDs of FinTech (i.e., artificial intelligence, blockchain, cloud computing, and big data), may very well transform the financial services industry.

In this report, we focus on AI and big data, which are at the forefront of the technological developments taking place in investment management. Blockchain and cloud computing tend to be closer to the technology infrastructure than the application layer and are beyond the scope of this investigation.

Since the publication of the 2018 FinTech report, our conversations with senior technology executives and researchers at some major financial firms and research institutions around the world left us with the impression that AI and big data applications in finance and investments have come off the "peak of inflated expectations" and are going through the "trough of disillusionment."[2]

To help investment professionals currently exploring different AI and big data applications and those who would like to start exploring applications of these technologies but do not know where or how to start, we set out to identify representative cases in AI and big data.

The aim of this research is to identify relevant use cases such that investment practitioners and firms can take appropriate actions now to navigate the changing landscape and prepare for investment success as AI technology reshapes the investment management industry.

What Do We Cover in This Report?

In *Fintech 2018: The Asia Pacific Edition*, we discussed at length what changes the ABCDs of FinTech—in particular, artificial intelligence—may enable in transforming the financial services industry. In *Investment Professional of the Future*,[3] we discussed key roles and skills for future investment teams against the backdrop of technological transformation. In this report, we extend both lines of thought to describe what some investment organizations are already doing to incorporate artificial intelligence and big data into their investment processes and how they are organized to make the change happen. Specifically, for each case study, we focus on three areas:

1. What changes are taking place in the investment process?
2. What artificial intelligence and big data technologies enabled these changes?
3. What are the key roles (and related skills required) in the team, and how do teams collaborate to effect change?

We believe the best way to achieve these objectives is by finding actual and concrete cases of AI and big data technology application that are live in production at credible investment organizations around the world. In selecting the cases presented in this report, we used the following criteria:

- Geography—cases from firms spanning three regions (i.e., the Americas, Asia Pacific, and Europe)
- Specialism—coverage across four major areas of the investment business: equities, debt, asset allocation, and hedge funds (alternatives)
- Practical application—only cases that are live in production, no proofs of concept

The cases also include both discretionary and systematic strategies. Our findings confirm a view that ML techniques are more suited to systematic strategies and big data is used more by discretionary managers.

1 Larry Cao, *Fintech 2018: The Asia Pacific Edition* (Charlottesville, VA: CFA Institute, 2018).
2 Gartner, Gartner Hype Cycle. www.gartner.com/en/research/methodologies/gartner-hype-cycle.
3 CFA Institute, *Investment Professional of the Future* (Charlottesville, VA: CFA Institute, 2019). https://futureprofessional.cfainstitute.org/.

We also included two cases from outside the "core" investment management area—namely, Goldman Sachs (focusing on its sell-side research business) and Ping An, a Chinese financial conglomerate that originated from the insurance business. Amid industry changes such as MiFID II, the value of sell-side research is increasingly coming into question. How can sell-side analysts demonstrate the power of their research? The Goldman Sachs case may provide some food for thought for investment managers. Ping An's case is powerful not only in the scale and tremendous growth of its AI team but also in the widespread penetration of AI and big data applications throughout its core businesses.

In addition, there are two special focus cases that drill deeper into specific aspects of AI and big data's application in investments. The case by Marcos López de Prado discusses how machine learning techniques can help improve the performance and stability of the MPT portfolios. The case by Yang Yongzhi describes how NLP and computer vision techniques can be used in gathering and processing vast amounts of information.

Overall, the cases cover operations ranging from a small team at an investment firm to massive technology teams at financial conglomerates. We hope this diversity will enable readers from firms of all sizes to relate to the findings.

We selected cases targeting diversity in

- Geography
- Specialism
- Practical application
- Size of firm/team
- AI/big data sophistication

APPLYING AI AND BIG DATA IN INVESTMENTS: CHALLENGES AND OPPORTUNITIES

AI and Big Data Application in Investments: The New Frontier

In the words of numerous industry heavyweights, AI is the new electricity. The *Economist* magazine championed the slogan, "Data is the new oil."[4] We discuss in this section how some firms are taking advantage of frontier powers in investment management.

AI: NLP, Computer Vision, and Voice Recognition

Researchers have made tremendous strides in building the ultimate "seeing, hearing, and understanding" machine in recent years.[5] In the case of natural language processing (NLP), computer vision, and voice recognition programs, AI is used to capture text, audio, and imagery from a variety of public sources and internal/vendor databases. Examples include transcribing analyst conference calls and extracting data from issuer filings for valuation models. In most cases, the program automates what is traditionally a manual and repetitive task performed by an analyst. We expect to see these types of applications being used more and more in the industry; they broaden the investment professionals' reach and improve efficiency by combing through multiple data sources and combining them into one platform.

Such programs also increase human capacity by freeing time otherwise spent on manual work. Junior analysts used to spend much of their research time finding and entering information. These routine and repetitive activities will likely become the first to be taken over by AI programs, which have a natural advantage in this type of work.

AI: Machine Learning and Deep Learning

More sophisticated programs will further process the information harvested from various sources to generate signals to inform the investment decision-making process. This often requires sophisticated AI techniques, such as ML and DL.

Machine learning is a general term for computing methods and algorithms that allow machines to uncover patterns without explicit programming instructions.[6] ML programs inform themselves how to interpret inputs and predict outputs.[7] Deep learning is a type of ML that is based on artificial neural networks (a type of learning modeled on the human brain).

DL algorithms are often applied to improve the results of NLP, computer vision, and voice recognition programs. They can also help extract useful information from large piles of data. For example, these algorithms can infer certain key words from conference call transcripts or identify sentiment from unstructured data, such as social media. Such information can then be translated into trading signals or, more simply, alerts for human analysts and portfolio managers to process.

ML and DL programs are also popular with quantitative (systematic) managers who often find it helpful to apply these techniques in order to improve the effectiveness of their quantitative processes. There are several cases in the report that illustrate this point.

Traditional statistics and econometrics are based on techniques first developed a couple of centuries ago, and their applications in finance often involve linear regression models. These linear models are effective in many situations. However, at least some of the complexities in the real world may be better captured using ML techniques because of their ability to handle contextual and nonlinear relationships, which can often arise in finance. For example, ML techniques may be more effective than linear regressions in the presence of multicollinearity (where explanatory variables are correlated).[8] In these cases, ML and DL techniques provide investment managers with additional toolkits that can give them an edge.[9]

4 See www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.
5 See Larry Cao, "Artificial Intelligence, Machine Learning, and Deep Learning: A Primer," *Enterprising Investor* blog (13 February 2018): https://blogs.cfainstitute.org/investor/2018/02/13/artificial-intelligence-machine-learning-and-deep-learning-in-investment-management-a-primer/.
6 K.C. Rasekhschaffe and R.C. Jones, "Machine Learning for Stock Selection," *Financial Analysts Journal*, vol. 75, no. 3 (Third Quarter 2019).
7 Cao, "Artificial Intelligence, Machine Learning, and Deep Learning" (2018).
8 See Rasekhschaffe and Jones (2019).
9 For suggested further reading on machine learning, including types of ML algorithms and their applications, see the CFA Institute Refresher Reading on "Machine Learning" (2020 Curriculum), available to CFA Institute members and charterholders at www.cfainstitute.org/en/membership/professional-development/refresher-readings/2020/machine-learning.

Big Data: Alternative Data and Unstructured Data

Data scientists define big data with four Vs: volume, variety, veracity, and velocity.[10] The terms often used in the investment circles are "alternative data" or "unstructured data."

Alternative data refers to data from sources that are not currently used or not yet mainstream. In comparison to structured data, which is data that are digitized and stored in relational databases, unstructured data refers to data that are often in text, image, or voice formats and are not readily processable. Alternative data and unstructured data are related and yet not quite the same. Alternative data is often unstructured when first discovered, and unstructured data is usually not used by mainstream investors, making it alternative.

Examples of alternative data/unstructured data often used in the investment world today include satellite images, earnings conference call recordings and transcripts, social media postings, consumer credit and debit card data, and e-commerce transactions.

Finding the new data source that generates alpha has become the next arms race among some analysts and investment managers, much like how managers have traditionally competed to find the unturned stone in the public markets. In a way, extracting signals from big data is simply an extension of what analysts used to do—visiting stores to check out customer traffic, for example. Now some of them use satellite images or sensor information of the parking lot to infer the same information. The new techniques offer efficiency gains; an analyst can cover a lot more stores in much less time using satellite imagery or sensor data.

Alternative data tends to be niche and is more popular with fundamental managers running discretionary portfolios who use these data as one input in the investment decision-making process. Some of the cases included in this report provide real-world examples of how alternative and unstructured data are being used.

Types of AI/Big Data Applications in Investments

- **AI: NLP, computer vision, and voice recognition.** Used to process text, image, and audio data.
- **AI: Machine learning.** Used to improve the effectiveness of algorithms used in investment processes.
- **Big data: Alternative and unstructured data.** Used to process alternative and unstructured data for investment insights.

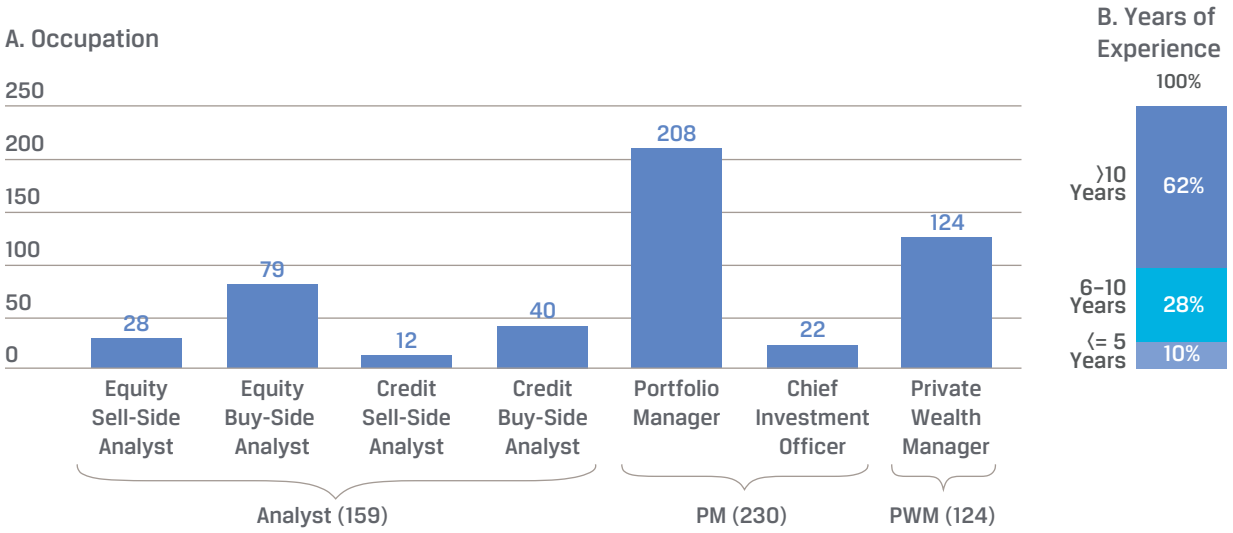
Where Does the Industry Stand in terms of Applying AI and Big Data?

CFA Institute conducted a practice analysis survey to understand the state of adoption of different technologies in the workflows of analysts, portfolio managers, and private wealth managers. This section provides some of the pertinent findings to illustrate the industry landscape regarding AI technologies and to set in context the specific case studies that follow.

The survey was sent to a randomized sample of CFA Institute charterholder members in March 2019, and there were a total of 734 respondents (52% from the Americas, 18% from Asia Pacific, 30% from Europe, Middle East and Africa). Respondent occupations spanned equity and credit analysis, portfolio management, chief investment officers, and private wealth management, as shown in Panel A of **Figure 1.** The professional experience of the respondents is shown in Panel B.

10 See www.ibmbigdatahub.com/infographic/four-vs-big-data.

FIGURE 1. SURVEY DEMOGRAPHICS



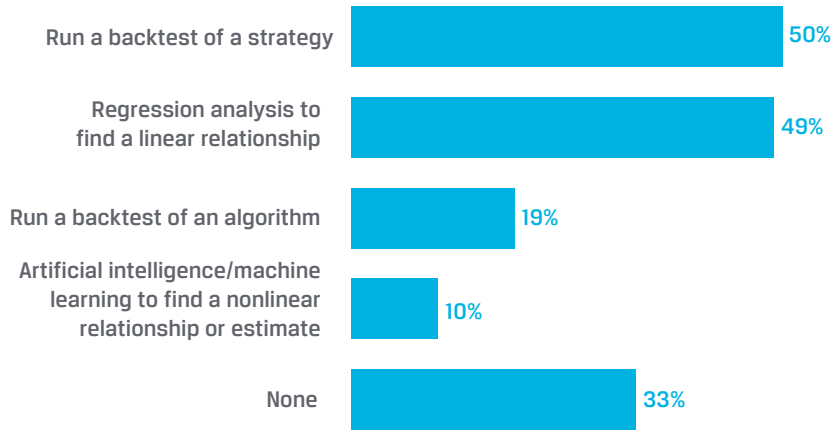
Note: Survey participation (N = 513).

The survey results indicate that few investment professionals are currently using programs typically utilized in ML techniques, including coding languages such as Python, R, and MATLAB. Most portfolio managers continue to rely on Excel (indicated by 95% of portfolio manager respondents) and desktop market data tools (three quarters of portfolio manager respondents) for their investment strategy and processes.

Moreover, as the results in **Figure 2** illustrate, only 10% of portfolio manager respondents have used AI/ML techniques in the past 12 months, and the number of respondents using linear regression in investment strategy and process outnumbers those using AI/ML techniques by almost five to one.

FIGURE 2. STATISTICAL TECHNIQUES USED IN INVESTMENT STRATEGY AND PROCESS

Portfolio Manager: Which of these have you used in the past 12 months for investment strategy and process?

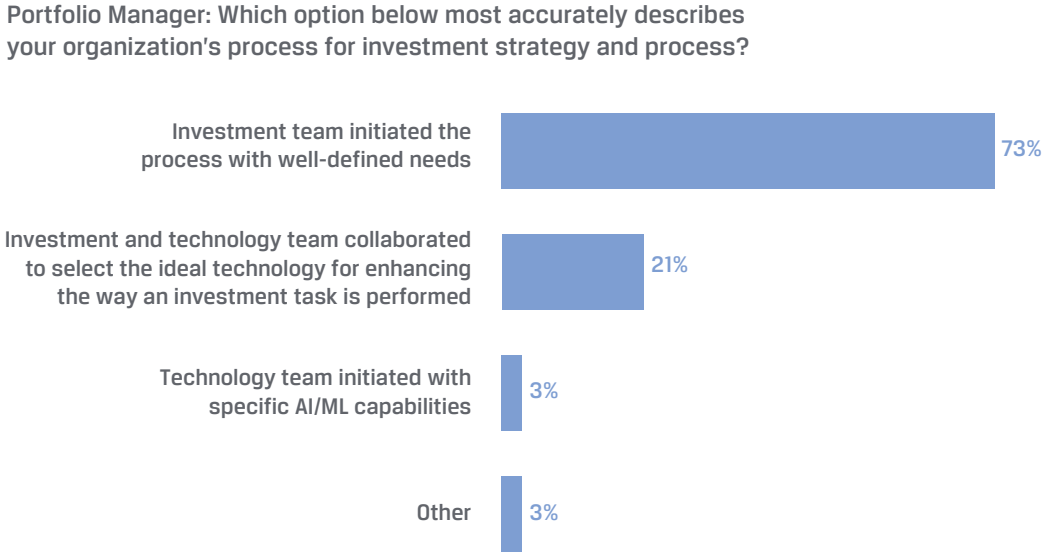


Note: Survey participation (N = 230).

Only 10% of the portfolio managers who responded to the survey used AI/ML techniques to improve their investment process in the past 12 months.

At the organizational level, the extent of collaboration between investment and technology teams remains relatively low, as shown in **Figure 3.** This suggests further integration may be needed to realize process efficiencies as these technologies take hold.

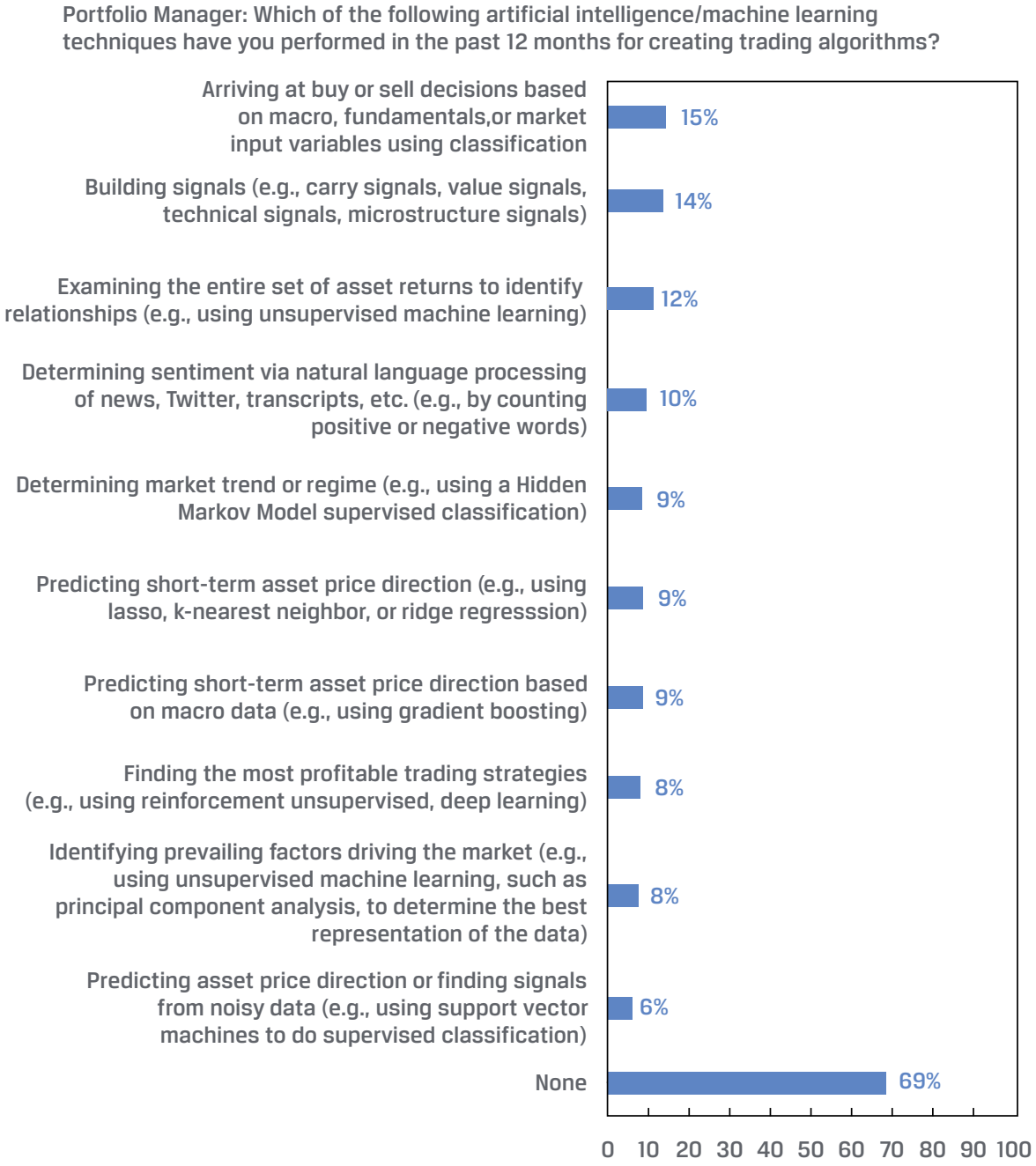
FIGURE 3. ORGANIZATIONAL RESPONSIBILITIES FOR INVESTMENT STRATEGY AND PROCESS



Note: Survey participation (N = 230).

The prevalence of AI/ML techniques in trading strategies is also low, according to the survey. As shown in **Figure 4,** 69% of portfolio manager respondents report not using any AI/ML techniques for creating trading algorithms in the past 12 months. Those professionals who are using these techniques indicate a wide range of use cases, including arriving at buy or sell decisions based on various input variables (15%), building signals (14%), and determining sentiment based on NLP (10%), among several others.

FIGURE 4. AI/ML TECHNIQUES USED FOR CREATING TRADING ALGORITHMS



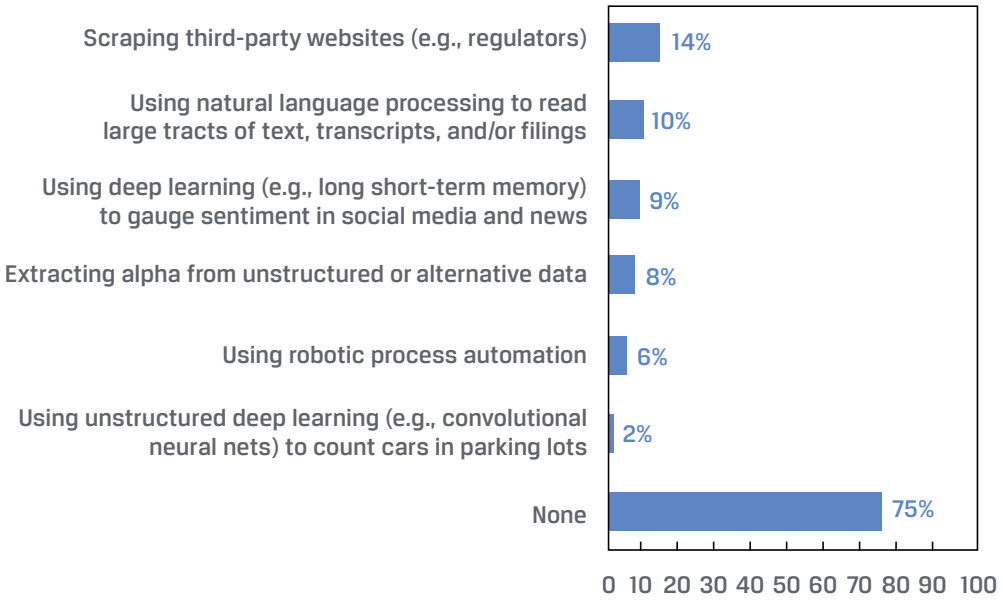
Note: Survey participation (N = 230).

A similar result emerges in Panel A of **Figure 5**, which shows that three-quarters of analyst respondents are not using AI/ML techniques for industry and company analysis. Of those who are, the two most popular techniques cited are scraping third-party websites (cited by 14% of respondents) and using NLP (cited by 10% of respondents). In comparison, 40% of respondents cited using linear regression for industry and company analysis (not shown).

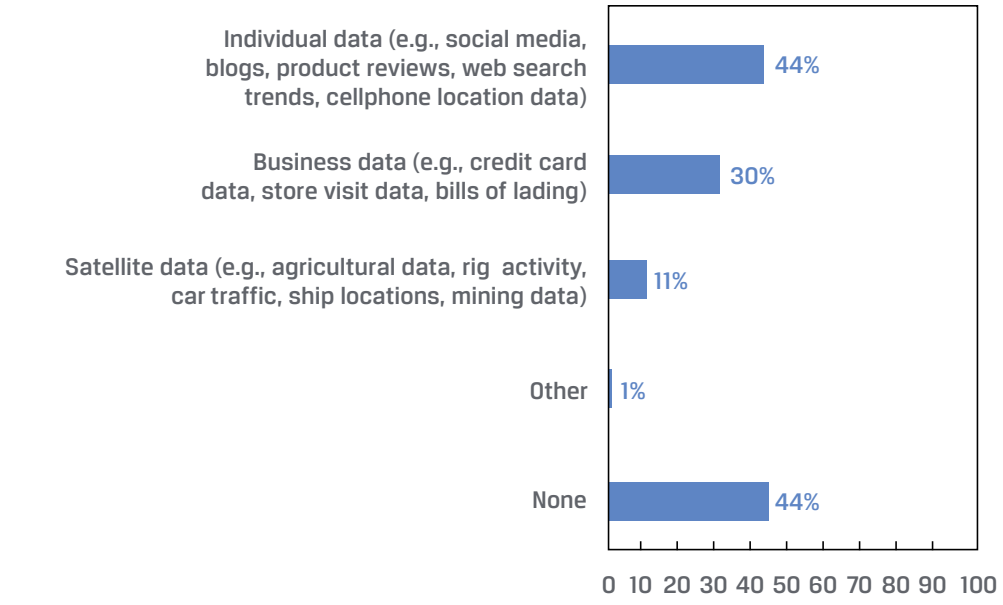
Using unstructured and alternative data for industry and company analysis is more popular than using AI/ML techniques among investment professionals. As illustrated in Panel B of Figure 5, 44% of analyst respondents report using individual data, such as social media, product reviews, and web search trends, in the past 12 months while only 11% have used satellite imagery. One caveat of these results, however, is that they do not allow us to infer how often or how extensively these data sources are being used in industry and company analysis. A significant number of professionals, 44%, report not using these data.

FIGURE 5. AI/ML TECHNIQUES VS. UNSTRUCTURED/ALTERNATIVE DATA FOR INDUSTRY AND COMPANY ANALYSIS

A. Analyst: Which of the following artificial intelligence/machine learning use cases have you performed in the past 12 months for industry and company analysis?



B. Analyst: What type(s) of unstructured and/or alternative data have you used for your industry and company analyses in the past 12 months?



Note: Survey participation (N = 159).

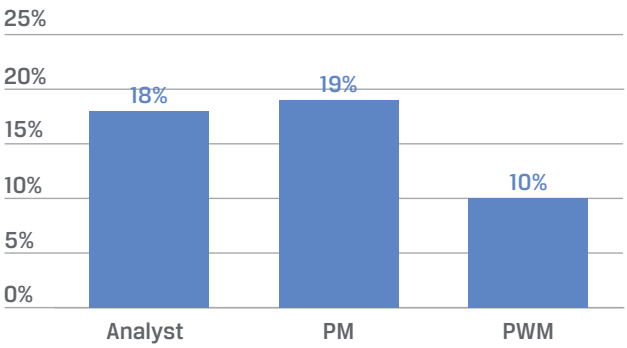
In sum, these results suggest that the investment industry is in the very early stages of adoption of AI techniques and related technologies, and few professionals are currently using AI/big data techniques in their daily investment processes.

However, approximately one fifth of analysts and portfolio managers report participating in AI/big data training, as illustrated in **Figure 6.**

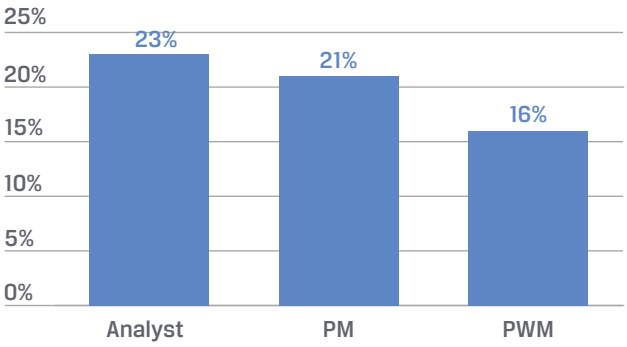
Overall, given the low current utilization of AI and big data techniques coupled with the large number of practitioners undergoing training in these fields, the industry seems poised to undergo significant growth in the coming years.

FIGURE 6. AI/ML VS. DATA ANALYTICS TRAINING

A. AI/ML Training Done in the Last 12 Months

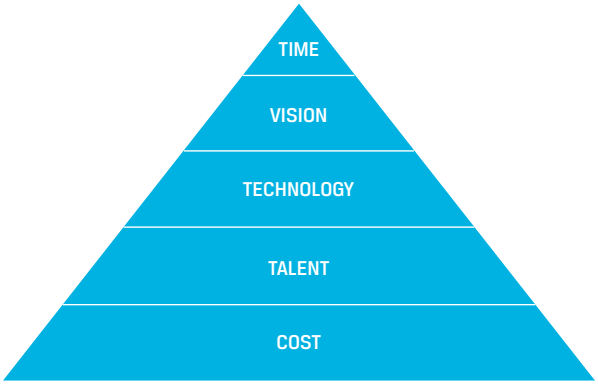


B. Data Analytics Training Done in the Last 12 Months



Challenges in Applying AI and Big Data: The FinTech Pyramid

FIGURE 7. THE FINTECH PYRAMID



So, what is holding investment professionals and investment firms back from realizing the full power of AI and big data? We have identified five major hurdles, which are described next in increasing order of difficulty.

Hurdle #1: Cost

Financial institutions are not strangers to huge IT budgets, but launching AI and big data capability can involve significant upfront cost as well as ongoing maintenance costs.

The high cost can at least in part be attributed to the new data sets that enable these technologies and that have been catching the industry's attention.[11] Identifying, cleansing, and making sense of these data sets is no small feat, which is why one prominent economist believes that small firms will find it increasingly difficult to compete in the age of AI and big data.[12]

Hurdle #2: Talent

College graduates with basic programming and statistics training, not to mention those with advanced degrees in AI or related fields, are already very popular with employers in the age of AI. Yet this is only part of the story.

There is a real advantage of working at one of the top technology companies that employs and invests significantly in AI. Google, Microsoft, Baidu, and Alibaba are some of the names that come to mind. Much of the latest and greatest developments in AI are taking place at these companies, and the small number of employees involved in these projects have become a rare breed who have access to knowledge and skills not currently taught yet in the top schools around the world. What further complicates matters is that it seems very few of the top AI talents are actually looking to work in the investment industry. Maybe AlphaGo and driverless cars are innately more exciting for AI scientists. Either way, obtaining talent appears at least a degree harder than managing the cost.

Hurdle #3: Technology

We are at the beginning of the AI revolution, and technology is still fast evolving. This creates significant challenges for those investing in AI applications because the risk of being leapfrogged by a latecomer is significant. Staying current with the latest developments is a real challenge for most investment professionals and organizations, barring a privileged few. Having a sizeable budget and top talent are prerequisites to staying ahead of the pack.

Similarly, in the alternative data space, the exploration of new data sources is in its relative infancy. There are many new data vendors entering the field, and extracting useful signals from the avalanche of data remains challenging.

Hurdle #4: Vision

There will likely be sweeping changes in the investment industry driven by advances in AI and big data technologies in the coming decades. These technological changes have to be managed from the top of organizations for them to fully penetrate the business while efficiently deploying resources.

As noted in *Investment Firm of the Future*, IT deployment in investment firms has been substantially reactive to date, with firms trying to marshal technology to capture efficiencies in the face of legacy issues.[13] Firms will need to focus on proactively developing the skills and procuring the systems to stay competitive. Strategic vision, leadership commitment, and collective ownership of IT deployment will be essential for firms to succeed in the future.

11 See, for example, www.cnbc.com/2017/11/28/making-millions-from-the-data-hidden-in-plain-sight.html.
12 www.nytimes.com/2018/01/12/business/ai-investing-humans-dominating.html.
13 CFA Institute, *Investment Firm of the Future* (Charlottesville, VA: CFA Institute 2018): 10–11.

Hurdle #5: Time

Any progress, no matter how small, often takes a significant investment of time, among other things. This is simply a fact of life when you are on the frontier of development.

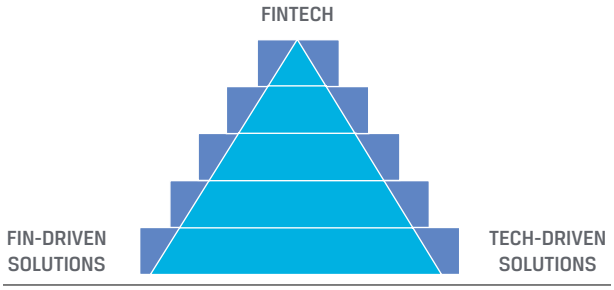
Every firm wants to be the first at turning over a rock and uncovering useful information, but exploring ways to increase alpha and integrating the new approaches into existing investment processes take time. Even in the most advanced markets and at firms where the most sophisticated technologies have been in deployment for many years, most big data projects still require a lot of time and effort to prepare the data and make them fit for purpose. Patience and persistence are necessary, and even then, many projects will not succeed. Time remains one of the toughest challenges to overcome, and success does not happen overnight.

Current State of Play and the Road Map to the Top of the Pyramid

Investment firms will need to substantially overcome the five hurdles (i.e., use the latest AI and big data technologies to solve core investment problems) to reach the top of the pyramid, where Fin meets Tech. But ascent to the top requires a collaborative approach; overcoming each hurdle requires consideration of both Fin and Tech dimensions.

Conceptually, this is illustrated in **Figure 8.** In the Fin corner, investment solutions tend to be driven by quants who come from a finance background. By and large, these types of solutions rely on existing data sets and do not rely heavily on alternative data, saving time and effort in identifying relevance (separating signal from noise) and testing and cleansing data. At the same time, they may not benefit from new information from alternative data and the latest technical breakthroughs in NLP, computer vision, and voice recognition, for example.

FIGURE 8. WHERE FIN MEETS TECH



Also in the Fin corner are some discretionary managers experimenting with new data sources that they come across. What is often missing is the overall strategy of systematically leveraging new technologies to gather and process new information that will feed into the investment process, creating an edge.

In the Tech corner, solutions are typically driven by technologists coming from outside the investment world. The architects and their teams tend to have in-depth knowledge of the latest AI and big data technologies and can create the fanciest wizardry that leverages the latest technology. Often, they are not built with a specific business objective or end user profile in mind and cannot be easily incorporated into the investment process of an established investment firm.

Introducing AI and big data into investments may be the single most significant change to the investment process that investment professionals will experience in their careers. Given the complexity illustrated by the FinTech pyramid, it will take many iterations to get everything right and reach the top of the pyramid. The important takeaway is the need to take a collaborative approach and expect to ascend the pyramid step by step. There is no shortcut.

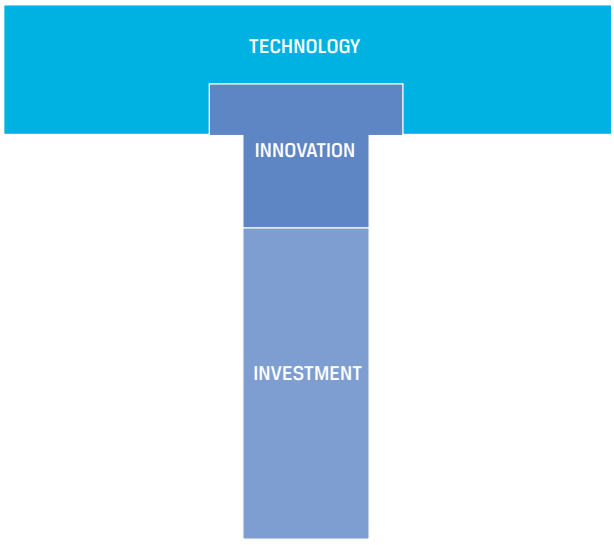
Making It Happen: T-Shaped Teams

The FinTech pyramid highlights the strategic imperatives of applying AI and big data in investments. The T-shaped team concept we introduce in this section provides an operational and organizational approach to making it happen.

We discussed in *Investment Professional of the Future*[14] the increasing importance for individual investment professionals to acquire T-shaped skills. T-shaped professionals have both domain-specific specialist knowledge and wider professional connections, understanding, and organizational perspective. In addition, T-shaped teams have a broad and deep collective intelligence and benefit from a collaborative culture and cognitive diversity.

In the context of AI and big data in investments, we can apply the concept of T-shaped teams. The combination of skills and collective intelligence gathered through investment expertise and technology application across investment strategies or processes is an example of a T-shaped team in this area. We also emphasize a third aspect of T-shaped teams in this context—namely, the role of innovators in connecting investment and technology teams, requiring such professionals to be particularly strong in T-shaped skills. The small T in the overall T-shape shown in **Figure 9** illustrates this aspect.

FIGURE 9. T-SHAPED TEAM



Under this framework, roles in the investment function are not substantially different from what we observe in the industry today, but that may not be true for the technology function. The technology function in future investment teams will likely require different skill sets than those required today. In particular, data scientists, in addition to computer engineers, will become important.

The third function, innovation, is of critical importance, because its main function is to facilitate collaboration between the investment and technology functions, something that the industry has not had a strong track record in to date. Innovators may have the title of researcher, strategist, product manager, or business developer. The lack of appreciation for this function is evidenced by the fact that professionals serving the roles often sit in different departments at different firms. Some sit in the investment or technology departments but must have a keen understanding of their counterparts' business to be effective collaborators. Strategists and product managers, for example, may take on the role because they have a better understanding of the big picture than the specialists from the investment and technology functions.

As noted in "FinTech and the Future of Financial Services," powerful FinTech will be the result of collaboration between powerful Fin (ancial institutions) and powerful Tech (nology companies).[15] The argument is true not only from a strategic perspective but also from an operational and organizational perspective.

In the early stage of collaboration, T-shaped teams are typically small in scale and often exist only on an informal, project-specific basis. As operations mature, T-shaped teams become more commonplace and are more permanent features of the organizational structure. The complexity of the issues at hand requires organizational commitment, which can be best identified by the number (and effectiveness) of the T-shaped teams an organization supports.

14 CFA Institute, *Investment Professional of the Future* (Charlottesville, VA: CFA Institute, 2019).
15 Larry Cao, "FinTech and the Future of Financial Services," in *Fintech 2017: China, Asia, and Beyond* (Charlottesville, VA: CFA Institute 2017). Available in Chinese at www1.hkej.com/dailynews/investment/article/1313726/.

Outlook and Word of Caution

Perspectives on Humans vs. Machines in Investment Management

Will AI and robots become so smart that they will replace us? The fear has lingered in the human psyche enough to pervade the current FinTech discussions.[16]

We believe in the power of the "AI + HI" model—that is, most tasks are and will remain best handled using both AI and human intelligence, and the collective power of the two is superior to either element on its own.[17] The path of adoption begins with routine, rudimentary tasks such as capturing information from texts and images, producing reports, and populating spreadsheet models, where AI has some advantage over human beings in the breadth of information they can process at high speeds. Analysts are then free for higher-value tasks that require more experience and judgment.

It is not a race between humans and machines. The competition ultimately is among "AI + HI" teams, and the stronger teams that effectively harness and combine both elements will outlast the weaker ones. The successful investment teams of the future will excel in collective intelligence through cognitive diversity (artificial and human) and T-shaped skills.

Perspectives on Applying AI and Big Data in Investment Management

Despite the important role they will play in the investment industry, AI and big data are no panacea. In some situations, additional information (big data) can add alpha, and in others, enhanced algorithms (ML) may detect previously undiscovered patterns. Still, AI and big data certainly won't provide all the answers investors need or want.

One of the challenges ML techniques face, for example, is that they work better in the test environment (i.e., based on the training data set) and may not always respond appropriately to new situations in the real world.

This is the problem of overfitting—where algorithms perform well in sample but poorly out of sample. AI may work for AlphaGo, where all the rules are set. The ever-changing investment world, however, presents more difficulties. In addition, at least some of the ML programs for business are more like a black box; users do not have access to the logic behind ML actions. As a result, some of the features captured by the programs have no causal relationship with the variables the models try to predict.

As technology and understanding progress, these challenges may be overcome, but as of now, we should put the power of AI and big data in perspective when embarking on a journey to explore the unknown.

Outlook

Based on our research, including interviews and conversations with academics and practitioners in both investments and technology, we can conclude that

1. AI and big data have the potential to bring about the most significant change to the investment management industry that current professionals will experience in their careers.
2. Successful investment firms of the future will start to strategically plan their integration of AI and big data techniques into their investment processes now.
3. Successful investment professionals will understand and exploit the opportunities brought about by these new technologies and applications, enabled by collaborative organizational cultures, cognitive diversity, and T-shaped teams.

The cases presented herein are instructive for investment firms and individuals as they adapt to the changing investment landscape. The experiences of these pioneering firms can provide a guidepost for the investment industry as it forges ahead in this new era of AI.

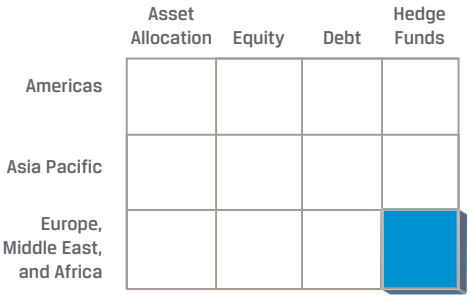
16 See, for example, www.forbes.com/sites/quora/2017/11/02/will-robots-eventually-replace-humans-as-the-dominant-species-on-earth/ and www.barrons.com/articles/will-ai-help-human-investors-or-replace-them-1543071600.

17 See Larry Cao, *Fintech 2018: The Asia Pacific Edition* (Charlottesville, VA: CFA Institute, 2018) and CFA Institute, *Investment Professional of the Future* (Charlottesville, VA: CFA Institute, 2019).

CASE STUDIES

1. ENHANCING TRADING STRATEGY AND EXECUTION WITH MACHINE LEARNING: MAN AHL

Contributor: Anthony Ledford[18]



Background

Man Group is an active investment management firm that provides long-only and alternative investment products with USD114.4 billion assets under management as of 30 June 2019. Man AHL is a quantitative investment manager that is part of Man Group; it is headquartered in London with USD29.9 billion assets under management as of 30 June 2019.

Man AHL trades a wide range of hedge fund and long-only investment strategies that typically span asset classes and geographies. All of these feature some degree of machine learning. For example, the execution process—common to all of Man AHL's strategies—uses an adaptive intelligent routing algorithm to pick the optimum route to market for a given trade. Machine learning is most widely used within our multi-strategy programs, where it is used in various guises, such as pattern recognition, trend following, and natural language processing.

Man AHL first started researching ML and its application in investments in 2009, but that effort did not lead to implementation in the program portfolio. In 2012, we approached it again, but this time, we used a different ML approach. The team was very skeptical, but after an extended period of research testing, paper trading, and live trading using Man AHL's own capital,[19] the first ML strategy eventually entered Man AHL's program portfolio in 2014.

Investment Process

So far, machine learning has had the most impact in two areas of Man AHL's investment management activity: first, in the development of trading strategies (these are the algorithms that generate trades—i.e., what to buy or sell and when to do so), and second, in improving the efficiency of execution of such trades (i.e., their delivery to and completion within financial markets). When we apply ML techniques while developing trading strategies, we put a lot of effort into ensuring that the resulting strategies diversify our existing portfolios. In other words, the ML algorithm does not perform an undirected search but instead is modified to seek features that provide both alpha and diversification and to discount or ignore existing portfolio features or nuisance effects we don't want to capture. Building up this research experience was hard, especially in the beginning, but we believe it is now a well-trodden path within the team.

Once the trading strategies have determined the current optimal positions (e.g., long 700 shares in ABC, short 250 shares in XYZ), any trades needed to obtain these positions are passed to the execution system. The job of the execution system is to trade to each desired position within a specified time horizon while minimizing the total transaction cost (market impact) incurred in doing so. Man AHL has been using its own electronic trade execution algorithms for over a decade, and they are complemented by third-party algorithms developed by broker or bank counterparties. The key question that needs to be answered—potentially thousands of times a day—is, What is the most efficient (cheapest) way to get each trade completed? That is, which algorithm should we use, or should a human deal with it instead?

This question is answered online, in real time, using ML techniques to automatically allocate trades to one of the available execution channels (e.g., Algo A, Algo B, ..., human). In our experience, human decision makers find this optimal routing problem notoriously difficult because of the nonstationarity and high degree of noise in transaction cost data. We started looking at this problem with tools from reinforcement learning at the end of 2016 and found benefits in both money and time saved. From modest beginnings trading a handful of futures contracts in Man AHL, the approach has been scaled up and is now used to help optimize the futures and cash equities order flow from across Man Group.

18 Chief Scientist, Man AHL. This case is prepared based on our interview with him and materials submitted by Man AHL.
19 Monitoring a new strategy in live trading using Man AHL's own capital is one of the final validation steps required before the strategy can enter the program portfolio.

AI/Big Data Technology

We are open-minded about the full spectrum of ML techniques; for example, we are agnostic as to whether we solve a problem with deep learning or Bayesian machine learning. Our job, I believe, is to explore the space, find out what works and what doesn't, and integrate the things that add value into client trading. In terms of strategies that have made it through the R&D phase and into the client portfolio, we have live trading with Bayesian ML, DL, and pattern recognition algorithms. More recently, strategies based on NLP, where the underlying data are text articles, are also now live in client trading. Researching such strategies requires specialist hardware—that is, a processor called a graphical processing unit (GPU) can complete the calculations required for DL research in 1/30th of the time taken by the CPU in a standard computer.

For researching trade execution strategies, the use of simple time-series data, even if sampled at high frequency, are inadequate because the relevant data object is more complicated—namely, the dynamically changing limit order book (LOB). Such market microstructure research can require orders-of-magnitude more data throughput and computing power than simple time-series data, especially when combined with ML models, which themselves can be computationally expensive, even for modestly sized data sets.

The choice of programming language is especially important. Man AHL has used Python for researching both trading and execution algorithms since 2011, and the live implementation of these algorithms for trading the client portfolio is also undertaken in Python. Previously, researchers used a combination of S-Plus, Matlab, and R, while live implementation was in C++ and Java. Using one language for both research and implementation has improved throughput—for example, by removing the need for double-coding—but has also assisted in less expected ways (e.g., enabling researchers and technologists to work closer together, on a single codebase).

Team Structure and Development Process

The great majority of the researchers at Man AHL have scientific backgrounds (e.g., in mathematics, statistics, computing, physics, engineering, or seismology, to name but a few). Members of our ML team have very diverse non-financial backgrounds (e.g., computational neuroscience, online news) and bring post-doctoral research experience to the application of ML techniques in different areas. Initially, we applied machine learning in futures markets because this area was where we had the highest familiarity. Later, once we'd learned some lessons about what worked and what did not, we broadened our research to cover cash equities. The chief investment officer and other senior members of Man AHL identify the main directions; the ML team and the rest of research then seek to capture trading signals within those areas, which are additive to our existing strategies. All staff are encouraged to complement this top-down perspective by suggesting research projects where they suspect a gap in our current systems or some feature that can be improved.

The teams are highly integrated. There is no clear distinction between researchers, technologists, and portfolio managers; they are all on a continuum. Successful researchers whose strategies have made it through to the client portfolio maintain oversight of the live trading and performance of their strategies.

Key Takeaways

This stuff is hard. Nothing works off the shelf. It is a common misperception that by putting a lot of data, fancy computers, and smart people together, you'll be able to extract useful signals. Experience matters. Only a fleetingly small percentage of data is "useful." Alternative data provides new opportunities but often does not have enough history.

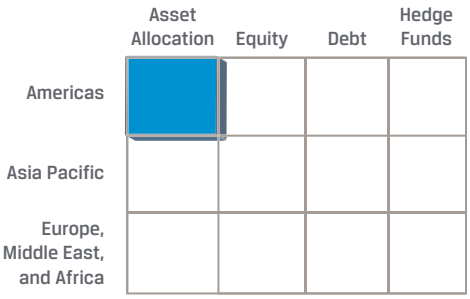
What you can extract using non-ML tools should be the benchmark for deciding whether to use machine learning. Is what extra you capture with machine learning worth the extra complexity? Our maxim is, therefore, "Use the simplest tool that does the job."

Embrace open source. Stay involved over the long term by contributing back. Form a virtuous circle.

Be bold in the process. Have resolve to decide what is worth pursuing, and kill off projects that don't look promising.

2. GENERATING SIGNALS FOR QUANT MODELS WITH MACHINE LEARNING: NEW YORK LIFE INVESTMENTS

Contributor: Poul Kristensen, CFA[20]



Background

New York Life Investments is the asset management arm of New York Life Insurance Company and had USD314 billion in AUM as of 31 December 2018.

The Multi-Asset Solutions team manages USD10 billion in global macro asset allocation products. The team members have a diverse set of backgrounds, spanning from economics and finance to engineering and physics.

Investment Process

We use a mix of fundamental and quantitative research inputs. Our quantitative toolbox serves as the starting point for our decision process. But with global macro investing, some market drivers are very hard to quantify, and a fundamental/qualitative assessment of the outlook is thus always important.

In the quantitative toolbox, we use a factor-driven approach to asset allocation, in which asset classes are decomposed into their underlying risk factors. We use predictive analytics for these risk factors that are based on four different systematic approaches: cycle, value, momentum, and sentiment. Within these quantitative approaches, we use ML techniques as important inputs. Our approach is designed, and the factors chosen, based on extensive simulation and cross-validation work as well as practical experience.

The model serves as the starting point for the discussion at our monthly investment policy meetings. AI tools are also used on an intra-month basis, when large shifts in markets and data flow occur. The tools are used by both portfolio managers and quantitative researchers supporting the portfolio managers.

The economic cycle can be an important driver of market returns, especially when major shifts occur in terms of acceleration or deceleration of the economy or outright downturns and recoveries. Such shifts in the cycle affect not only the mean of returns on risky assets but also the wider characteristics of their probability distributions, such as volatility, skewness (asymmetry), and kurtosis ("fat" tails). ML techniques are of great help in mapping the economic cycle and "nowcasting" the market regime based on a wide range of incoming data on a day-to-day basis. ML techniques allow us to incorporate far more data—a wider range of indicators—than traditional statistical techniques could handle. This enables us to get a more reliable real-time read on cycle trends.

AI/Big Data Technology

Predictive analytics for the current regime of the cycle (e.g., up and down phases) are used as an input in cycle investing. Predictive analytics for the speculative-grade default rate are used as an input when assessing value signals for credit-sensitive asset classes.

Some indicators are the same as before ML techniques were introduced, but we are now able to use and process—and thus condense the signal contained within—a much wider range of indicators. ML techniques allow us to process the signal contained in a large number of indicators more reliably and faster, thus enhancing the workflow process.

AI and ML techniques are not only applied to portfolio decision making itself, but they are also increasingly affecting how we work with data in a broader sense. AI techniques can be used to map and rank predictors of particular events by their importance and set up dashboards monitoring the most critical indicators in an efficient manner. The workflow involved in creating such dashboards is far more efficient with AI and ML techniques.

Development Process

The quantitative toolbox, including ML tools, was developed and is maintained mostly by three to four team members. These team members have backgrounds in econometrics, physics and data science, operations research, and engineering.

The inspiration was university working papers documenting the benefits of ML techniques in investment strategies. Based on ideas from such working papers, as well as insights from practitioners

20 Managing Director, Economist, and Portfolio Manager, New York Life Investments. This case is prepared based on our interview with him and materials submitted by New York Life Investments.

within the firm, the team has engaged in projects exploring the use of AI for investment decisions in different contexts in collaboration with investment teams across the organization. Developing innovative techniques is hard; the reality is that there is no textbook answer. We are exploring different use cases and experimenting with different ideas. And as with all innovation, "trial and error" is a natural part of the process. Some techniques show promise early on, while others do not deliver results or are too complex to explain and map to real decisions. Then the researchers return to the drawing board. It's an iterative process, and we expect that to continue going forward. There is no free lunch.

The key "aha" moments often arise when exploring the data, experimenting with an idea, and reading the output of the model. Suddenly, an idea comes to mind. Data visualization is thus key to the creative process. In addition, sometimes "aha" moments can arise when going through working papers and notes by researchers using new techniques. Even though the use cases may not be exactly equivalent, there may sometimes be useful insights.

Impact/Key Takeaways

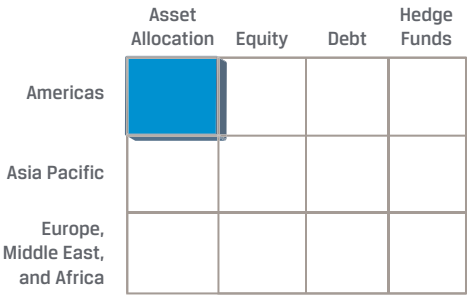
We use a large number of indicators of credit quality, leverage, liquidity, borrowing and issuance trends, delinquencies, and so forth as inputs in predictive tools for credit spreads and defaults. ML techniques have enabled us to incorporate larger volumes of data, improve the accuracy of the predictions, and identify the most important predictors to monitor in dashboards.

The signals generated from ML techniques help us focus on the most important indicators when making portfolio decisions and thus better assess risks and opportunities in markets.

The signals generated from ML techniques—especially the cycle and value signals—help us focus on the most important indicators when making portfolio decisions and thus better assess risks and opportunities in markets. The cycle framework has also allowed us to monitor a wider range of indicators when gauging a portfolio's outlook, saving time in our preparations for investment committee meetings as well as in the daily monitoring of portfolios.

3. REFINING EQUITY TRADING VOLUME PREDICTION WITH DEEP LEARNING: STATE STREET CORPORATION

Contributor: Dajun Wang, CFA[21]



Background

Trading volume prediction is an important topic in finance, especially for those institutions providing asset management or trading execution services. It is important because it not only helps an institution better allocate its trading resources but also gives traders a better view of market conditions. In this trading project, we explore the idea of predicting trading volume with DL technologies and package that process into a service capable of forecasting daily equity trading volumes segmented by markets.

Inputs

A financial institution may serve funds of many investment styles, so the temporal patterns of its daily trading behavior are fairly complex and must be separated by a number of factors, including the following:

1. Financial market indexes. This kind of factor is especially important for active funds. The investment strategy of an actively managed fund usually relies on a set of financial market indexes, such as the CBOE Volatility Index or Merrill Lynch MOVE Index, as inputs. Thus, the movements of these indexes affect the fund's daily trading volume.
2. Market index reconstitution schedule. This kind of factor is likely to affect passive funds. A passively managed fund usually tracks a market index, such as the S&P 500 Index or Hang Seng Index. A reconstitution on these indexes generally results in higher trading volume on the effective day. This phenomenon is widely observed.
3. Historical trading volumes. Historical trading volumes provide the baseline for normal trading levels and the long-term trend.
4. Regional market temporal correlations. Some financial institutions operate around the globe, which includes multiple regional markets. Even though each market operates on its own local holidays and schedules, correlations are still observed among these markets under certain scenarios.
5. Special calendar days. Some calendar days (e.g., Thursday, the last business day in a month, public holidays) have significant influence on daily trading volume.

To provide a predictive service with accuracy that meets a comprehensive financial service provider's requirements, our model has to take all of these factors into consideration. The service ingests these factors daily and returns the predicted trading volume by markets. When the input factors' data accumulate to a certain size, the model retrains itself with the updated data to accommodate potential structural changes.

AI Technology

The core of our predictive service is a sequence-to-sequence model based on a multi-layer convolutional neural network (CNN). The model is able to take in several time series of arbitrary length and output predicted series with the last element as the final prediction.

The model contains two encoder networks and one decoder network. The first encoder applies proper transformation on market index data and extracts features from them. The second encoder does the same to calendar indicators and historical trading volumes. The output of the two encoders is combined and then fed into the decoder network. The decoder network is built with an "attention" mechanism. It contains both CNN layers and dilated CNN layers. The CNN layers in the decoder network are mainly designed to capture short-term information from features, and the dilated CNN layers are designed to capture long-term and periodical information.

The DL framework of Facebook's PyTorch was used to build the model. The implementation is capable of using both CPU and GPU for model training and calculation.

Team Structure and Development Process

Our model development team has two data scientists, one financial engineer, and one business analyst. The team members have education backgrounds that cover mathematics, statistics, machine learning, and quantitative finance.

21 Managing Director, State Street Corporation. This case is prepared based on our interview with him and materials submitted by State Street Corporation.

The project was initiated in August 2018. Since then, the model development team has actively worked together to understand the business background and requirements, discuss potential features and data sources, address data issues, and confirm intermediate results. Review meetings with management occurred monthly. Model developers provided updates and working examples at these meetings to stakeholders, who in turn provided the modeling team with ideas for model enhancements.

When the input factors' data accumulate to a certain size, the model retrains itself with the updated data to accommodate potential structural changes.

In November 2018, a structure of the model was fixed. The modeling team began to wrap the model into a service. In this phase, the model validation team and IT team were involved. The model validation team has one internal model validator who verifies the model's performance under various scenarios. He also advises the model development team on model robustness. The IT team has one architect, two developers, and one quality assurance engineer. They refactor the complete model into a RESTful service in a secure and resilient way. After that, users can access our service with multiple program languages, such as C/C++, Java, Python, Matlab, or with whatever HTTP clients choose, such as internet browsers. Their contributions also include user authentication and authorization, resilient deployment of the model, encrypted data transactions, disaster recovery, and audit logging.

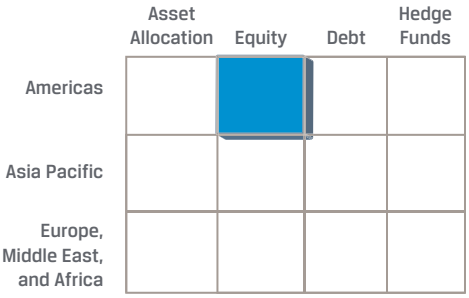
In January 2019, the first beta version of the service was released and tested with State Street internal trading data. The test data set includes daily equity trading volume data covering 63 markets and funds of various trading styles. During the test, real data are fed into our application daily to verify its functionality and prediction accuracy. Our model reached an R2 of 0.86 for the US market and 0.54 for other markets during the test, which is as expected and comparable with the matrix from the training data set. Our service has also successfully foreseen several trade volume spikes triggered by FTSE index reconstitution in the test.

Key Takeaways

Our trading volume prediction project proves that AI technology can be a powerful addition to State Street's front-office core investment functions. It establishes a standard for future projects on converting AI-assisted models into practical and capable tools that drive the bottom line. Future projects will rely on this experience to further improve the company's operation while complying with standards on architecture design, data governance, and project life cycle management.

4. LEVERAGING AI/ALTERNATIVE DATA ANALYSIS IN SELL-SIDE RESEARCH: GOLDMAN SACHS

Contributors: Ingrid Tierens, PhD, CFA[22]
Dan Duggan, PhD[23]



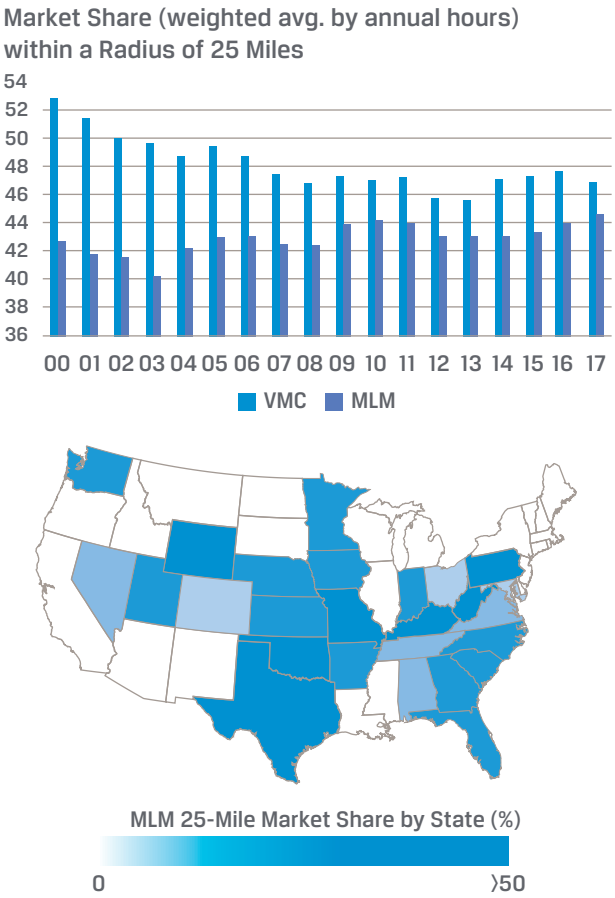
Background

Goldman Sachs' Global Investment Research Data Strategy team works directly with both equity and macro research analysts around the globe on projects that require analytical and quantitative skill sets, with the research analysts providing the subject matter expertise and spelling out the investment use cases. Over the past two-plus years, the team (based in New York City and Bengaluru, India) have collaborated on nearly 200 published research analyses across various sectors and markets worldwide. Data sets include app analytics, social media, satellite imagery, government-sourced data, and newspaper articles. Combining these sources with a variety of advanced techniques, such as clustering or sentiment analysis, can create unique investment research insights. One example is "GS Aggregates Share Tracker: Geospatial Assessment Points to Pricing Upside," published by lead equity analyst Jerry Revich, CFA, on 17 January 2019, which leverages publicly available geospatial data to create local market share estimates and map it to local demographics.

Analytical Approach

The GS Aggregates Share Tracker answers the following questions relevant to a subsector of the materials industry: (1) how to understand positioning in a hyperlocal industry, (2) how to represent market share for public companies in a largely private competitive landscape, and (3) how to inform investment professionals about the directional sense of quarterly company results with respect to aggregate volumes. In the aggregates industry (the mining of sand, gravel, and crushed rock for use in concrete for the construction industry), volume growth and product pricing margins are two of the most important key performance indicators. The markets themselves are hyperlocal, meaning that market share calculated at the 25-mile range is more accurate than those at the state or even the county level. Identifying where a company's quarries are located as well as providing a quarry-level estimate of production volume deepen the analyst's understanding of how a company is positioned (e.g., in which markets a company has pricing power or material exposure). The addition of a new data source that meaningfully tracks company performance with geospatial information provides better insight into how these companies operate. **Figure 10**, which reproduces Exhibit 21 and Exhibit 8 in Revich's 2019 report, visualizes the added insight for Martin Marietta (MLM) by showing its 25-mile market share by state and comparing it with the market share of Vulcan Materials (VMC) over time.

FIGURE 10. MARTIN MARIETTA MARKET SHARE BY STATE AND COMPARISON WITH VULCAN MATERIALS MARKET SHARE



Sources: Mine Safety and Health Administration, company data, and Goldman Sachs Global Investment Research.

22 Managing Director, Global Investment Research, Goldman Sachs. This case is prepared based on our interview with her and materials submitted by Goldman Sachs.

23 Vice President, Global Investment Research, Goldman Sachs.

Estimating aggregates volumes in this space can inform a company's strategy: which markets are in focus, how potential M&A activity may change the company's positioning, and where pricing margins are most important. At its most basic level, the analysis is used to estimate organic growth on a quarterly basis and to provide directional input ahead of earnings. In addition, the analysis is used to understand market share shifts and pricing impact and is levered to estimate company-level exposure weighted by aggregate volumes to weather events, such as regional flooding and hurricanes.

AI/Alternative Data Analysis

The GS Aggregates Share Tracker combines both company data and publicly available quarry data as inputs to the model. It takes advantage of open-source geospatial libraries and leverages the quarry specific metadata to provide estimates of both aggregates production and location information. An iterative validation sequence allows corrections to the raw data, which reduces uncertainties associated with the derived locations. Entity mapping is performed through a combination of NLP and company-reported subsidiary data, in conjunction with the analysts' domain expertise. Market share calculations are location specific and thus are calculated for each of the roughly 9,000 US quarries at distances from 5 to 50 miles in 5-mile increments.

Team Structure and Development Process

To achieve our objective of offering buy-side clients better and more differentiated investment insights using alternative data and machine learning, we aim for the intersection of the following three functions: (1) tap into domain expertise, (2) access all relevant information, including both "traditional" sources and alternative data where appropriate, and (3) apply advanced analysis techniques to extract relevant insights.

Whether the project is situation specific or longer term, it is absolutely critical to first understand the investment research use case we are trying to solve. Otherwise, we could end up wasting a lot of time and effort. Communication between the research domain experts and the quants is key, and the more opportunities to understand the research process, the better.

Conversations with research analysts help us determine whether our value added as a team comes from embedding alternative data, providing data analysis expertise, or helping analysts systematize their insights into a data product. Sometimes we discover it is not the right project for us to be involved in.

The model creation process of the quant research team is fully integrated with the analyst team, and both teams are involved with building and validation from start to finish. For the GS Aggregates Share Tracker, our quant team worked with the covering analyst team led by Jerry Revich on every step of the process: iterating entity mapping results for each company and cross-checking company quarry location model mapping, historical aggregate production results, and specific validation sequences based on significant events on a per company basis. The initial models relied only on public company data, but as the model has grown in complexity, it has expanded to cover the entire 9,000+ US quarry industry, creating tremendous scope expansion for the analyst team.

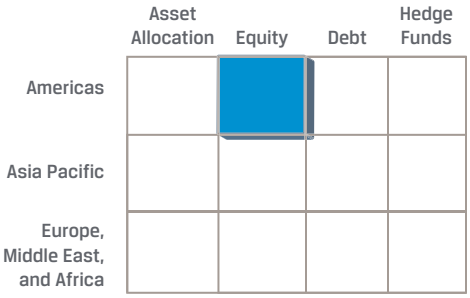
Key Takeaways

Some advice based on our experience:

- Don't underestimate the potential of more advanced techniques and approaches for making investment professionals more efficient and thus freeing them up to spend more time on creating better alpha insights!
- Alternative data does not necessarily translate into an advantage for systematic managers. In fact, we think that more niche, sector-specific data sets lend themselves much better to a fundamental analyst or portfolio manager who digs deeply into specific assets as opposed to an analyst or portfolio manager who covers a much broader range of assets.
- In our experience, a single data set or a single methodology has never been the sole driver of a research product. For us, alternative data adds to the mosaic, but it is not a goal in and of itself.
- Good research analysts are often aware of data that is relevant to the assets they cover, but they may not know how to best access and/or analyze the data.
- Leveraging alternative data doesn't necessarily mean breaking the bank. There is quite a bit of publicly available data that can be additive when you properly combine it with other data sources you may already be looking into.

5. DISSECTING EARNINGS CONFERENCE CALLS WITH AI AND BIG DATA: AMERICAN CENTURY

Contributor: Tal Sansani[24]



Background

American Century Investments was founded in 1958 in Kansas City. The firm now manages over USD160 billion of assets for financial intermediaries, institutional clients, and individual investors. The Disciplined Equity team comprises research, portfolio management, and investment technology professionals.

The bulk of work to integrate AI techniques into the investment process has taken place within the Disciplined Equity team, which is based in the heart of Silicon Valley. About USD14 billion of assets are managed by the team using processes that incorporate elements of AI and big data technologies.

Investment Process

American Century's stock-selection model is a direct input into the rebalancing process, alongside a slew of risk factors, transaction cost factors, and many other considerations. Portfolio managers review trade lists with a keen eye for name-level risks, emerging macro news, and changes to the model outputs.

The stock-selection model seeks to identify fundamental drivers of future returns in a systematic fashion. It is built on four pillars: quality, growth, valuation, and sentiment. Each of these pillars is composed of roughly five proprietary factors. The model predictions are continuously updated as new information becomes available.

The sentiment pillar is designed to capture fundamental changes to a company's prospects using information that isn't fully reflected in its latest financial statements. For example, our models glean information from short sellers, identifying companies that are beneficially exposed to trending product markets and geographies while systematically avoiding stocks that show signs of crowding/herding.

A key component of our sentiment model analyzes the language coming from management in its quarterly conference calls. Our conference call deception model is made up of four components: omission (failure to disclose key details), spin (exaggeration from management and overly scripted language), obfuscation (overly complicated storytelling), and blame (deflection of responsibility). To avoid biases, we account for both the unique style of a given management team and the collective language of its industry peers.

AI/Big Data Technology

Our machine-reading models are heavily reliant on NLP. Within that domain, we leverage a variety of techniques in sentiment analysis, document classification, part-of-speech tagging, entity recognition, and topic detection. Many of these language-processing techniques require machine learning to perform accurately and efficiently.

Although less glamorous, there are also a variety of technologies required to scrape, clean, and structure messy/unstructured data. For example, data on the web require HTML/XML parsing tools, string pattern recognition to key in on specific content, and systemized removal of noise (e.g., advertisements or irrelevant links). More so than ever, the need to parse multiple languages is also part of the process. The tools required range from techniques that have existed for decades (e.g., regular expressions and string distance functions) to rapidly evolving open-source platforms in Python and R to powerful and targeted services from third-party technology companies.

The larger the dataset and more pre-processing required, the more computational resources are needed. Although our team commonly leverages a locally managed, high-performance computing cluster, larger tasks require the elastic resourcing of cloud services like Amazon Web Services or Google Cloud.

24 Vice President and Senior Quantitative Researcher, American Century. This case is prepared based on our interview with him and materials submitted by American Century.

Team Structure and Development Process

The team's best ideas—using AI or not—often start with dialogue and input from portfolio managers and researchers, working together to generate unique, economically sensible stock-selection ideas. Our goal is to translate fundamental insights into systematic, objective, and repeatable algorithms.

Once the idea is qualitatively vetted, researchers work with investment technologists (often one researcher and one technologist) to source and process the necessary data inputs and collaborate on the necessary technologies and techniques.

In this particular case, I led the research and development process and served dual roles on the team, as a researcher and an investment technologist. However, many individuals on the team were consulted during the process: The chief investment officer, researchers, and portfolio managers all provided feedback and insight throughout the R&D process. Technologists served as sounding boards for back-end design decisions.

Investment teams are most successful when working side by side with technologists toward a solution that is both economically sound and thoughtfully implemented.

My academic background is in applied mathematics and computer science, but most of my financial and technical training has been on-the-job learning—from colleagues and mentors and self-initiated.

The development process was highly iterative, from both a research and a technology perspective. I started with a baseline set of hypotheses, drawn from financial journals and psychology books, outlining the sorts of linguistic patterns associated with deception. As the research went along, I sought feedback from researchers and portfolio managers on the team—asking, for example, "Does this align with what you look for when you listen in on calls? How can the model be improved or fine-tuned?" One such enhancement involved contextualizing the model based on the industry in question and/or management's specific personality (e.g., a young, overly optimistic technology CEO vs. a long-tenured oil exploration CEO).

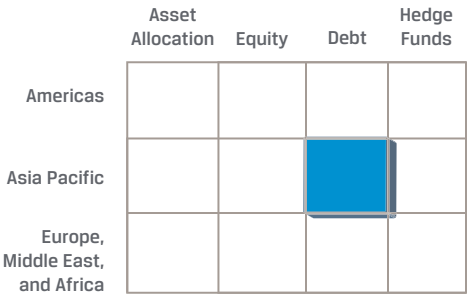
The development of this model required a multi-disciplinary research approach. This went so far as studying psychology textbooks to survey the signs of deception from a variety of contexts and people (e.g., children, criminals, business executives, etc.). The aha moment of this research was when we began seeing clear, shared linguistic patterns across deceivers in everyday life and polished company officials.

Key Takeaways

- *Focus on the investment problem first.* Our data scientists have a mix of technical backgrounds, with computer science and statistics being the most common. It's worth noting that while these backgrounds provide a great foundation, they are difficult to apply without domain knowledge in equity investing. Investment teams are most successful when working side by side with technologists toward a solution that is both economically sound and thoughtfully implemented.
- *A machine is only as intelligent as the data it learns from.* Sourcing additional data is often more effective than building a more elaborate model. The more comprehensive the training data, the more generalized the result of the machine processing new events, thereby mitigating common pitfalls like overfitting. Harnessing more varied data is especially critical to equity investing because the information relevant to valuations goes beyond traditional financial statements.
- *Humans and machines are complementary forces.* Human analysts are skilled at scrutinizing a small set of companies, and a robo-analyst can systematically apply its objective and unique insights across thousands of companies at a time. We believe teams that harness the benefits of both are well positioned in a highly competitive landscape.

6. AI AND BIG DATA ASSIST IN DEBT PORTFOLIO MANAGEMENT: CHINA LIFE ASSET MANAGEMENT AND CHINA SECURITIES CREDIT INVESTMENT

Contributors: Yang Rong[25]
Qu Jing[26]
Fan Siwen[27]
Tian Qiaomei[28]



Background

China Life Asset Management Company Limited (CLAMC) is China's largest asset management institution, involved in domestic and foreign public markets, public (mutual) funds, and alternative investments, among other fields. CLAMC manages AUM of more than CNY3 trillion (USD436 billion) as of the end of 2018, of which the majority is held in its fixed-income portfolio. Therefore, having a technology solution in place is particularly critical for CLAMC.

China Securities Credit Investment Co., Ltd. (CSCI), is a credit management technology service provider headquartered in Shenzhen, China. CreditMaster is an integrated credit management solution the company developed for the Chinese debt market. CLAMC started using CreditMaster in July 2018.

Methodology

Valued at USD13 trillion,[29] China's growing debt market has a relatively short history and faces various disclosure issues. "Non-standard products," such as investments in trust products, have grown to over CNY330 billion in CLAMC portfolios. CLAMC wanted a technology solution that would allow it to value fixed-income products in the investment process and to monitor developments of the broad market as well as individual issues.

CreditMaster offers various capabilities that satisfy CLAMC's needs:

- Credit risk data (i.e., datamart) includes data organized around entities, financial instruments, risk alerts, a credit evaluation engine, evaluation history, and user accounts/access rights.
- The evaluation engine incorporates ratings from rating agencies or other third parties and can generate a customized rating based on the previously mentioned data and other data in real time.
- The analytics module integrates CreditMaster into the credit analysts and risk managers' workflow. The module also allows clients to fine-tune model parameters based on their own analysis and experience.
- The automated report allows clients to automate standard credit reports with information available in CreditMaster, with analysts completing the report with their own conclusions.

The solution is used by credit analysts, fixed-income portfolio managers, and compliance teams at CLAMC in the credit analysis and investment monitoring process.

During this credit analysis and investment process, the system will automatically complete about 80% of the analysis based on external and internal data as well as models and ratings. Publicly available information includes such things as company annual reports/financial statements, legal proceedings, regulatory filings, and news—for example, information regarding over 20,000 equity and debt issuers and related financial institutions, as well as financial and operating information of 1,700 municipal and municipal corporate debt issuers.

CLAMC analysts and portfolio managers apply logit models when predicting the likelihood of default of any particular issue. The system will also recommend issues based on the risk and return profiles of the current and target portfolios. The system also generates alerts automatically. Portfolio managers can then decide based on their analysis what to do with a particular issue that is on alert.

25 Managing Director, Fixed Income, China Life Asset Management. This case is prepared based on our interviews with the contributors and materials submitted by China Life Asset Management and China Securities Credit Investment.
26 General Manager, Large Enterprises Credit Management Business Unit, China Securities Credit Investment.
27 Senior Vice President, Strategy and Development, China Securities Credit Investment.
28 Senior Product Manager, China Securities Credit Investment.
29 See www.cnbc.com/2019/04/01/china-bonds-debut-on-bloomberg-barclays-global-aggregate-index.html.

CreditMaster saves credit analysts' time and effort switching from system to system by integrating external and internal sources of information and modeling. At review sessions for issuer ratings, CLAMC analysts no longer have to print out data and documents from various sources; they can all be presented directly from the system. Doing so frees the analysts to focus their attention on conducting analysis and forming conclusions.

Similarly, risk managers no longer need to pore over multiple websites to track news and social media postings on the issuers. Data and news gathered from public websites and social media and purchased from third-party vendors are incorporated into the system, which generates alerts on issuers as it picks up signals.

AI/Big Data Technology

The system leverages AI and ML models to provide investment professionals with a tool that gathers and monitors comprehensive risk information, including:

- The distributed web spider system that can check tens of thousands of websites in a minute and collects millions of data points daily from news, bulletin boards, and blog sites as well as announcements;
- NLP that captures information from the dynamic unstructured text data collected for analysis and modeling (e.g., the fingerprint information generated with SimHash algorithms can help screen out same topic information and keep about 30% of the information that is relevant);
- The BiLSTM (bidirectional long short-term memory) model that performs named entity recognition analysis on issuers and products to determine whether the article focuses on them or simply mentions them;
- Text convolutional neural networks that perform sentiment analysis on text samples with over 90% accuracy when identifying negative information (which is very helpful for the alert system—it can send out relevant messages on holdings in the portfolio to the teams);
- The knowledge graph that enabled CSCI to establish a network that covers 70 million companies—hundreds of millions of equity ownership and corporate appointment data points—which allows the system to further leverage knowledge reasoning techniques and perform in-depth mining on related-party relationships (which is very helpful in preventing the transmission of risk when risk events happen at a related company of the monitored entity).

Team Structure and Development Process

The CSCI team serves multiple functions ranging from credit modeling to systems development. The main roles are as follows:

- *Credit analysts* develop and maintain the methodology of CSCI's industry credit models, which offer an alternative to clients' own credit models.
- *Quants* use quantitative and ML techniques to develop and maintain CSCI's industry credit models.
- *Data scientists* use various AI techniques, such as NLP and text convolutional neural networks, to collect and process data from public sources and third-party vendors to generate signals.
- *Engineers*, by far the largest component in terms of head count, are responsible for systems development work from the front end to the back end, including user interface and databases.
- *Product managers* communicate client needs to the engineers as well as quants and data scientists.

CSCI strongly prefers professionals who are not only strong in their own roles but can also branch out to one or two other functional areas.

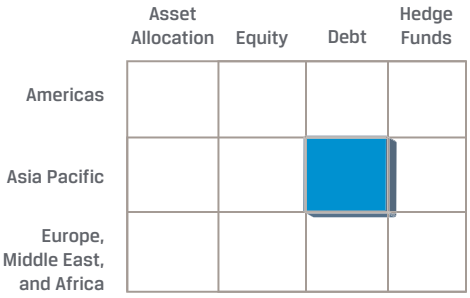
Key Takeaways

Developing an integrated system requires different skill sets, from the standardization and cleansing of different data sets to generating insights from those data sets. CSCI finds it essential to have the five roles centralized in one team to work toward one objective. The traditional model of separating data scientists, quants, and systems development into different departments does not work.

Between major releases that require sign-off from CLAMC management, the CSCI team delivers monthly "iterations" to the end users. The CSCI team finds the approach very effective in receiving client feedback in a timely fashion, and it reduces the risk of miscommunication between the investment and technology functions across organizations.

7. APPLYING AI AND BIG DATA TECHNOLOGIES IN THE FILING AND PROCESSING OF INSURANCE CLAIMS AND ASSESSING CORPORATE RISK: PING AN

Contributor: Xiao Jing[30]



Background

Established in 1988, Ping An Group is now the largest global insurance group with 1.8 million employees. Ping An is committed to being a technology-enabled retail financial services group, encompassing insurance, banking, and investments. Leveraging ABCDS (artificial intelligence, blockchain, cloud computing, big data, and information security) technologies, it has also built comprehensive capabilities in five ecosystems: financial services, health care, auto services, real estate services, and smart city services.

Ping An has always had an abundance of high-quality financial data. Ever since the AI team first started in 2015 with fewer than 10 people, we saw big data's potential in meeting users' personalized needs and helping Ping An grow its business. Four years later, the team is 1,000-strong, covering such business lines such as finance, health care, and smart city and working on various applications—for example, risk control, fraud detection, intelligent health care, medical diagnoses and treatment, operation optimization, intelligent finance, and precise marketing. Here we focus on the AI+ flash claim platform.

Business Process

Ping An SMART Flash Claim automatically completes the recognition of vehicle damage through image recognition, handling tens of thousands of cases claimed per day. Customers only need to take a few pictures of the damaged car and file the claim on the accident site, and the claim will be automatically processed in seconds with precise payment calculation. The system involves a series of key modules, including picture quality assessment, insured car verification, car part segmentation, damage identification for each part, payment calculation, and fraud detection.

The system has been running at Ping An for over a year, successfully processing over 30 thousand claims each day. It not only improves claim processing efficiency and thus customer experience but also stops potential frauds of billions of yuan. This system is now available to the insurance industry through the Ping An OneConnect platform. Ping An handles more than 11 million claims annually, with 98.7% paid within one day and 60% self-service.

Another application is in corporate risk assessment. There are over 70 million registered enterprises in China. Their information comes from three major sources: commercial registration and daily operations, public announcements including news and social media, and business relationships (e.g., supply chain, investment, legal, executives, and so on). To organize and analyze such rich and dynamically evolving data, Ping An developed OlaTop, an enterprise knowledge graph application.

Combined with advanced AI technologies, such as knowledge graphs, NLP, and machine learning, OlaTop precisely locates and solves the defects and intractable problems in traditional analysis framework of enterprise operation. It is fully developed in the field of financial analysis, producing mature tools, refined functions, and customized solutions, including such functions as financial analysis, bond analysis, sentiment analysis, industry chain analysis, and policy analysis. All functions can be implemented in four main areas: risk management, precision marketing, smart investment, and public service.

AI/Big Data Technology

A few AI techniques were developed to enable the AI+ flash claim platform's functionalities, such as image processing and segmentation, object detection and recognition, knowledge graph, and anomaly classification.

In the case of OlaTop, millions of legal proceedings are automatically interpreted and over 40 million lawsuit relationships have been extracted and incorporated into the graph. Signals regarding the enterprises from more than 300 news and social sites are automatically collected—hundreds of thousands of articles daily—and updated every 10 minutes. Deep graph analysis algorithms are then developed to support business decisions, such as risk assessment.

30 Chief Scientist, Ping An Group. He leads the Artificial Intelligence Center at Ping An Technologies. This case is prepared based on our interview with him and materials submitted by Ping An.

Business logics are directly integrated into OlaTop. For example, risk events are designed with different business logics for investments, bonds, and loans, and related signals are extracted through sentiment analysis from social and news data. Also, different orders of upstream and downstream relationships are encoded as risk indicators. When a risk event occurs upstream, the influence passes through the graph network and decays by the relationship orders.

Team Structure and Development Process

The artificial intelligence center of Ping An Technology is composed of business application teams and technology application teams. Among all the teams, two of them work together to deliver knowledge graph–based intelligent solutions: One is responsible for big data analysis, data modeling, general tool development, and so forth, and the other one is responsible for knowledge extraction, text analysis, and map construction. The former cleans and integrates the structured and unstructured data, and then the latter uses its knowledge extraction and graph construction capabilities to deliver the final solutions.

At present, the team has more than 1,000 artificial intelligence experts, mostly from top universities at home and abroad, such as Carnegie Mellon University, Harvard University, Peking University, Stanford University, Tsinghua University, the University of Oxford, and the University of Science and Technology of China. The core members of the team are technology leaders in the fields of computer vision, NLP, medical image processing, and data analysis, with work experience of over 20 years, on average. Before joining Ping An, they held important positions at top high-tech companies, such as Google, Microsoft, IBM, and Uber.

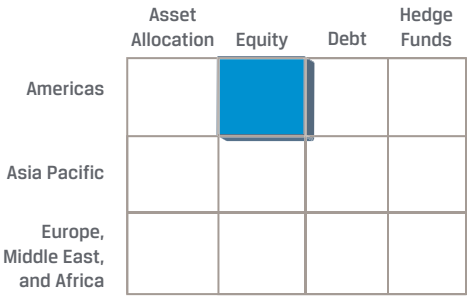
The system has been running at Ping An for over a year, successfully processing over 30 thousand claims each day. It not only improves claim processing efficiency and thus customer experience but also stops potential frauds of billions of yuan.

As for the development process, our inspirations usually come from a lot of pain points in the massive business, which drives business experts and AI scientists to find corresponding technical solutions through brainstorming. Each participant will contribute his or her own expertise and opinions, and eventually the team will reach a consensus and formulate implementation plans.

The development of Ping An's AI has gone through four stages, including the infant stage of building basic perception ability, the student stage of building knowledge graph systems, the expert stage of outputting intelligent solutions combined with business scenarios, and the creative stage of challenging the most difficult areas of human intelligence. At each stage, Ping An adopts an agile approach and, by setting clear objectives, work plans and priorities, we bring value to the business helped by rapid product iterations and regular reviews.

8. SENTIMENT ANALYSIS: BLOOMBERG

Contributor: Gary Kazantsev[31]



Background

Founded in 1981, Bloomberg provides financial software tools, such as securities trading platforms, data services, news, and analytics to financial companies and organizations through the Bloomberg Terminal and its Enterprise products and data feeds. The company started investing in ML and AI applications over a decade ago, ranging from improving the customer support experience to improving the Bloomberg Terminal user experience through natural language understanding and question answering, as well as information extraction and innovative news and analytics applications.

One of the earliest such applications was the sentiment analysis product, which has been available to customers since 2009. Since then, the product has grown and expanded from English to other languages (e.g., Japanese), from equities to commodities and currencies markets, and from news to other content types, such as social media (e.g., Twitter). The product can now also tackle other issues—named entity recognition and disambiguation, topic clustering, theme detection, market impact analysis, and others. While the application of machine learning in this context was novel back in 2009, it is now used by key players across the global capital markets for a variety of applications, from risk analysis and portfolio construction to alpha generation.

Today, this product is just one of a spectrum of advanced analytical tools powered by machine learning that Bloomberg offers.

Methodology

A variety of approaches to sentiment analysis have been explored in academic literature and in the industry. When we started development on this product in 2009, Bloomberg chose to approach it using cutting-edge technology. To this end, we used supervised machine learning and developed novel ensemble construction methods.

The key issue in text analysis problems is the selection of the target variable—that is, precisely formulating the question we wish the machine to answer. Many possible perspectives on this problem exist. For example, some seek to assess the internal state, or opinion, of the author of the document. There are applications where that perspective is of interest. However, in finance, this introduces additional ambiguities and complications to an already difficult problem. In contrast, we chose to predict the opinion of the reader—that is, a consumer of the news story who is a participant in the market. The question we chose to ask is, roughly, If you are a long-position investor in the underlying company and you read the story in absence of other information, would you take it as a positive, negative, or neutral event? This change in perspective, and thereby the target variable for the ML algorithm, makes a big difference in practice.

Training data for these models were collected in-house, and we trained annotators who had an appropriate background in finance. Special care had to be taken with respect to sampling to ensure that a sufficient diversity of sources, document lengths, topics, market caps, industries, financial periods, and so forth, were represented in the data. The key measure to inform model development is the inter-annotator agreement—that is, How often do human annotators agree on the answer to the posed question? If the agreement rate is low (i.e., no different from chance), the problem is ill posed. If the agreement rate is high, then there is a high likelihood we can build a machine to reliably reproduce this aspect of human judgment.

Using this carefully chosen question, we are able to obtain agreement rates between 80% and 92%—for news. For social media like Twitter, agreement rates for every type of language understanding task are generally lower. This number is naturally an intrinsic boundary to the accuracy of any classifier we build. It is important to note that simple methods, such as lexicons (dictionaries), are unable to solve this complex and ambiguous problem with the required degree of accuracy. We end up building much more complex models—nonlinear support vector machines, models incorporating complex linguistic features, and, most recently, recurrent neural networks—and integrating them into ensembles using data-driven methods. The resulting ML ensembles are real-time systems, processing a stream of more than 2 million

31 Head of Quant Technology Strategy, Bloomberg. This case is prepared based on our interview with him and materials submitted by Bloomberg.

documents per day—in just tens of milliseconds per document. The per story and entity sentiment data (e.g., for one document mentioning both Google and Microsoft, you may see positive sentiment for Google with 75% probability and negative for Microsoft with 50% probability) can then be aggregated into a confidence-weighted per-security time series, or index, to be used in portfolio construction, strategy development, or risk analysis.

Sentiment analysis data have been used by a wide variety of clients, both on the Bloomberg Terminal and in context of our Enterprise Event Driven Feed (EDF) product. On the Bloomberg Terminal, there are functions that allow clients to visualize the data, correlate them with other variables, set alerts on them, or use them to monitor a portfolio. The sentiment data are also used for automated story generation. In the Enterprise space, sentiment data have been used as a factor in portfolio construction and optimization, risk analysis, trading strategy development at various periodicities and for various holding periods, execution, and statistical arbitrage. Over the years, the composition of the consumers of sentiment data has changed. Initially, quant hedge funds and market makers were the primary users of this "alternative" data, whereas today, most institutional clients either use it or are looking to incorporate it into their workflows.

AI/Big Data Technology

Named entity recognition and disambiguation (NER/NED), topic classification, sentiment analysis, and clustering are all examples of the application of modern machine learning in the NLP domain. These tools can be used much more broadly across the industry to understand unstructured data, be that instant messaging chats, emails, analyst reports, quarterly filings and annual reports, compliance notes, or other trade-related information. For example, publicly available APIs from a variety of cloud vendors now make it possible to effectively process the audio stream of an earnings call into a text transcript. Once the data are in textual form, these same tools can be applied to build integrated market impact models that update as the call progresses.

Team Structure and Development Process

The overall system that supports and enables these advanced analytics is a large and complex technological solution that enables clients to perform many other workflows—search, alerts, research, and so on. Hundreds of people were involved in its development over the span of more than 10 years. At this point, the AI and Data Science Engineering organization at Bloomberg numbers over 200 people with advanced backgrounds in software engineering and high-performance distributed systems, mathematics, computational linguistics, machine learning, and natural language processing.

The development of these products is a very client-driven activity, and we work closely with customers to design and build solutions to these problems. For example, a small engineering team consisting of just two people started to build sentiment analysis in response to a client request. Having successfully demonstrated an initial prototype in the first six months of development, the team grew and expanded to deliver the product to clients and extend the technology to other aspects of the problem.

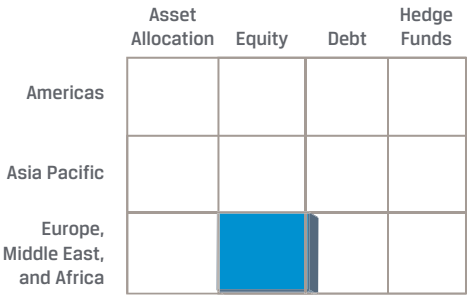
A decade ago, machine learning was exotic. The state of tools and libraries for computational linguistics, natural language understanding, and machine learning was much more basic. As a result, the majority of the development had to be undertaken in-house. The development process emphasized the need for a distributed computation environment and natural language understanding tools tuned to the finance domain, which spurred a major, multi-year investment in these areas. Given these investments in infrastructure and data processes (i.e., distributed computing, GPU hardware), the team can now execute these and similar projects much more rapidly, drastically minimizing our time to market with new solutions.

Key Takeaways

There were several key takeaways from the sentiment analysis product's development process and the broader text analysis initiative that followed. As previously alluded to, end-to-end evaluation is important. Careful selection of the target variable is critical. It is important to start with simple baselines and perform error analysis at every iteration of the model development; otherwise, we can easily find that a particular date or quarter is positive or negative by itself—because of inherent bias in the underlying data, for example. For this reason and others, interpretable models and human-in-the-loop are very important to us.

9. BUILDING THE DATA SCIENCE TEAM: SCHRODERS

Contributors: Ben Wicks[32]
Mark Ainsworth[33]



Background

The Data Insights Unit at Schroders was formed in 2014 by Ben Wicks and Mark Ainsworth. In discussion with the head of investment, Ben was asked to consider how Schroders could invest to gain an edge over its competitors in data-led research and gain early, differentiated insights into individual companies.

Use Case: Investment Idea Generation

After the release of a UN report on the theme of smart cities, a fund manager decided to search for companies that might benefit from this theme and increasing urbanization globally.

The first step was to search for articles that mentioned the phrase "smart cities" and the word "future," which yielded many thousands of matching articles. We then applied a number of ML techniques:

- Using a set of NLP algorithms, we extracted the sentiment, concepts, topics, entities, and keywords mentioned in the article. Crucially, we also identified any corporate entities being mentioned, including the "main company."
- The next step was to simplify further by using a kind of dimension reduction method. The particular algorithm we used operates using a force-directed graph, in which points representing the news articles are laid on a 2D plane, and it then iteratively "pulls" similar articles close to each other and pushes dissimilar articles away from each other. Over time, the result is that clusters of articles that talk about the same sort of ideas are gathered together.
- Once laid out in this way, the articles are assigned to clusters using an unsupervised ML algorithm.

The end result is a visual map, laying out a huge quantity of news articles onto the screen with similar articles grouped together and color-coded according to their assigned cluster.

The clusters were profiled by the name of the companies mentioned in the articles, the "main company" being attached to the map as text. A visual inspection of this "news map" showed a number of interesting features. In particular, there was a distinct cluster of articles at one edge of this map that mentioned a company the fund manager hadn't heard of before.

Having been tipped off to the potential for this small and little-covered company to be invested in by the fund, the fund manager conducted his financial analysis of its business and concluded that its fundamentals and prospects were good. Its shares were added to the portfolio, and roughly two years later, it was bought out at a +25% return to the fund.

Team Structure and Development Process

At the time, Ben was a portfolio manager on the Global and International Equities team. He believed that modern data science techniques and new data sources could indeed be used as a resource to complement existing, traditional stock market analysis to create a powerful combination that would invest for the long term, with an insight edge over peers.

The finance sector hasn't always kept up with data science and analytics in the way that academia and other sectors have, and thus, it was necessary to bring in a senior leader from outside of the industry. Mark Ainsworth has 20 years of data science and analytics experience from working in diverse companies/industries such as Formula One, telecommunications, and retail. Ben believed that these techniques could be used at an asset manager to give fundamental investors an edge.

Mark joined Schroders in 2014 and, with Ben, set about forming the team. After proving the potential with several successful proofs of concept, he started to bring in data consultants and data scientists to grow the team's capabilities. A team of data engineers from inside IT were brought together to help set up the infrastructure needed to pipeline the data while Mark, Ben, and the developing team set about bringing in new data sets and finding more use cases.

32 Head of Data Insights and Research Innovation, Schroders. This case is prepared based on our interview with the contributors and materials submitted by Schroders.
33 Head of Data Insights and Analytics, Schroders.

The data consultants and scientists who were hired came from a variety of backgrounds. Some were from commercial companies with large, well-built data science and analytics teams that worked on proprietary data in order to help make decisions about the direction of the companies. Others were from academic or scientific labs where data science skills were necessary. And from the start, the team included fresh talent straight out of university.

Development Process

Investors who were interested in the idea of a data science function also helped to influence the sort of data sets that were brought in and the kinds of projects that the team worked on. A crucial part of the process at this early stage was a series of regular brainstorming meetings between some of the investment professionals and the data scientists, which provided an invaluable opportunity to generate and test ideas and identify areas of shared interest among the multiple investment teams.

In the first few years, there were some valuable lessons about how to fit a team like this into a fundamental asset manager. Everyone in the firm was different. Some loved the idea of the team and what it would be capable of doing from the minute it was announced, but not all were as enthusiastic. What united everyone was their desire to see results. When valuable insights were delivered, there was even more positive feedback internally about what the team could do. This led to more investors who wanted to be involved in giving feedback on what data sets or types of work they would find useful.

All investment desks have used work from the Data Insights Unit—either from a piece of bespoke work for an analyst or from reports from the automated tools and regular report we built using large data sets and ML algorithms, which are delivered to their inboxes. The work is used by a diverse range of investors for many purposes, from analysis of individual stock names to the use of wide-ranging macro data.

AI and Big Data Technologies

A crucial element of the team's work has always been the technologies used; a good shorthand for the team's role is to provide insights from data sets that can't be analysed using Excel (after all, if it can be done in Excel, investment analysts are perfectly capable of doing it themselves). As a result, the team's work is centered around particular tools and techniques.

One tool is big data—huge data sets that are too big to be processed by a single computer, often with many billions of data points. For these data sets, the team uses a combination of cloud technologies (including massively parallel data warehouse tools like Amazon Redshift) and our on-premise Hadoop cluster using the Hive database.

Another tool is geospatial data, the kind of data that only makes sense placed on a map. For this, the team uses specialist tools like QGIS (a popular open-source geographical information system tool for viewing and generating maps) and PostGIS (a powerful database for holding and processing such data).

The team also has powerful techniques for harnessing the patterns and predictability in data, including machine learning and Bayesian inference. For these, we use powerful GPU-powered machines as well as Spark on our Hadoop cluster.

Day to day, the main languages used by the team are R and Python. We use Kubernetes and Docker to deploy the tools we've built into production. A very commonly used tool for making simple visualizations available to people is Tableau, although we also use R Shiny and Dash to build analytical dashboards. As always with such tools, the important task is to strip away distraction and deliver relevant information with clarity so that the user can see the insights.

Machine learning has been a powerful tool for us. We now have several ML-powered tools in regular use by our investment colleagues. One of the most effective uses of Bayesian methods is to identify when there have been significant changes in a time series that tracks brand sentiment, which we can then flag to our investment colleagues. The research analysts and fund managers can blend this information with everything else they know about the companies and their prices to judge whether they should make changes to their portfolio.

Key Takeaways

The two big lessons we have learned are (1) to have a senior sponsor who really believes in this sort of innovation for support and (2) to have the right mix of skills within the team. Experts from outside of the industry are necessary to build the team alongside a domain expert who knows what sort of insights are useful.

Building a data science capability sounds very exciting and is very trendy, but it is entirely pointless without positive, helpful outcomes. Some quick wins early on helped us to prove the value of the team. A data

science capability can't be an exercise in doing something fashionable for the sake of it; it needs to add value to the business, whether that's bringing in alternative data or building ML algorithms to look for significant changes in the pattern of a data set.

The team has always been very open about what we do and what we can do, and the decision was taken very early on to make the team a centralized function available to all investors. This meant that nobody worried about what using the data or the team's skills was costing them or if they were using a resource that they weren't meant to. Crucially, this decision also set up the team to find areas of value that spanned multiple investment teams and thus create things that no single investment team would be able to afford.

The signals from a large data set are best used to help create a picture of a company and/or the environment that it operates in rather than to launch a trade in isolation without considering other factors.

10. SPECIAL FOCUS: ENHANCING THE MPT EFFICIENT FRONTIER WITH MACHINE LEARNING

Contributor: Marcos López de Prado[34]

Background

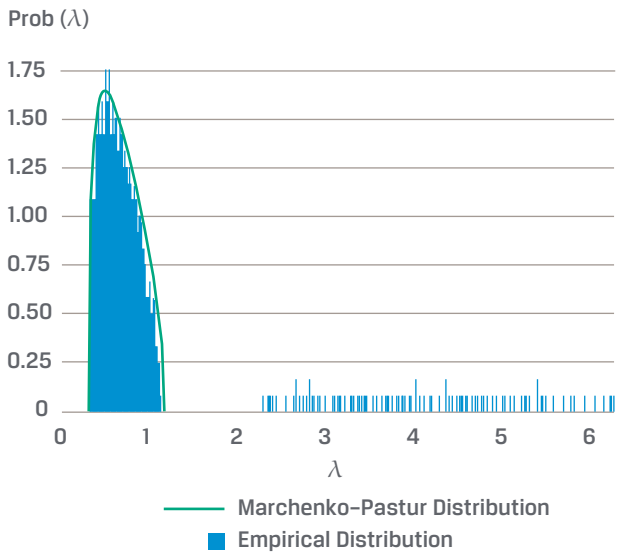
Since the 1950s, Markowitz's modern portfolio theory (MPT) has been the most widely used approach for asset allocation and has inspired many extensions, including the Black–Litterman approach.[35] In practice, these convex optimization solutions deliver poor performance, to the point of entirely offsetting the benefits of diversification.[36] For example, in the context of financial applications, it is known that portfolios optimized in-sample tend to underperform the naive (equal-weighted) allocation out-of-sample.[37] The explanation for this underperformance is that these optimal portfolios cannot be estimated robustly for two reasons: noise-induced instability and signal-induced instability. López de Prado spoke to us for this report about how machine learning techniques address the pitfalls of convex optimization in general and MPT in particular.

Noise-Induced Portfolio Instability

Financial correlation matrices are not robustly estimated. This is evidenced by that fact that even factor-based correlation matrices, such as those provided by BARRA, exhibit an extremely low signal-to-noise ratio. The reason is that we need at least ½ $N(N+1)$ independent observations to ensure that the covariance matrix is not ill-conditioned. For instance, a relatively small correlation matrix of only 100 instruments would require a minimum of 5,050 independent and identically distributed observations, or more than 20 years of daily data. Not only that, but correlations would have to remain constant for two decades, and returns would have to be drawn from a Gaussian process. Clearly, these are unrealistic assumptions.

To understand this problem, consider the distribution of eigenvalues plotted in **Figure 11**. The figure shows that over 90% of the eigenvalues of a factor-based financial correlation matrix fall under the Marchenko–Pastur distribution. The Marchenko-Pastur distribution predicts the distribution of eigenvalues from a random correlation matrix. In other words, almost all of the eigenvalues are associated with noise—even after applying factor models that impose a structure on the returns.

FIGURE 11. DISTRIBUTION OF EIGENVALUES COMPUTED ON A FACTOR-BASED CORRELATION MATRIX



To correct for this problem, Potter, Bouchaud, and Laloux recommend to de-noise the correlation matrix.[38] This procedure consists of discriminating between eigenvalues associated with noise and eigenvalues associated with signal with the help of the kernel density estimator (KDE), an ML algorithm. Eigenvalues associated with noise can then be centered by replacing each of their values with their average value. **Figure 12** plots the eigenvalues of the factor-based correlation matrix before and after de-noising. A critical difference between shrinkage[39] and de-noising is that the former averages all eigenvalues, including eigenvalues associated with signal, while the latter averages only eigenvalues associated with noise. In general, we should prefer de-noising over shrinkage, because de-noising helps stabilize the covariance matrix with minimal loss of signal.

34 CIO of True Positive Technologies, LP, professor of practice at Cornell University, and author of *Advances in Financial Machine Learning* (Wiley, 2018) and *Machine Learning for Financial Researchers* (Cambridge University Press, forthcoming 2019). Before launching his firm, he was a principal at AQR Capital Management and its first head of machine learning. This case is based on our interview with him and materials submitted by him.

35 H.M. Markowitz, "Portfolio Selection," *Journal of Finance* 7 (March 1952): 77–91 and F. Black and R. Litterman, "Asset Allocation: Combining Investor Views with Market Equilibrium," *Journal of Fixed Income* 1 (Fall 1991): 7–18.

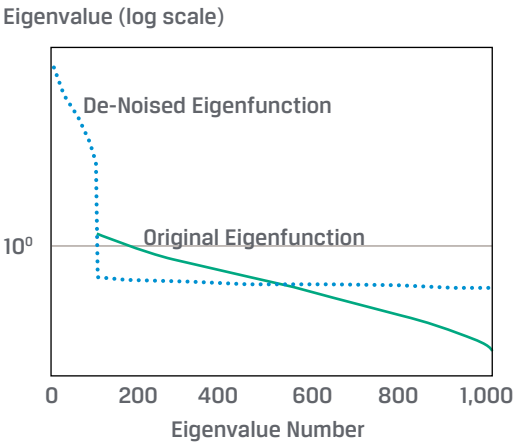
36 R. Michaud, *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation* (Oxford, UK: Oxford University Press, 1998).

37 V. DeMiguel, L. Garlappi, and R. Uppal, "Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" *Review of Financial Studies* 22 (May 2009): 1915–53.

38 M. Potter, J.-P. Bouchaud, and L. Laloux, "Financial Applications of Random Matrix Theory: Old Laces and New Pieces," *Acta Physica Polonica B* 36 (September 2005): 2767–84.

39 See, for example, O. Ledoit and M. Wolf, "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis* 88 (February 2004): 365–411.

FIGURE 12. DE-NOISING OF THE CORRELATION MATRIX



De-noising preserves the trace of the correlation matrix while reducing the matrix's condition number. The result is a correlation matrix with better numerical properties. We can demonstrate this through the following Monte Carlo experiment: (1) using a factor-based correlation matrix, draw a random empirical correlation matrix and vector of means; (2) de-noise the empirical correlation matrix; (3) derive the maximum Sharpe ratio portfolio from (2); and (4) compute the root-mean-squared error (RMSE) between the true optimal weights (derived from the factor-based correlation matrix) and the estimated optimal weights (derived from the de-noised empirical correlation matrix).

Table 1 reports the results from 1,000 Monte Carlo experiments, which show that de-noising is much more effective than shrinkage: The de-noised maximum Sharpe ratio portfolio incurs only 0.04% of the RMSE incurred by the maximum Sharpe ratio portfolio without de-noising. That is a 94.44% reduction in RMSE from de-noising alone, compared with a 70.77% reduction using Ledoit–Wolf shrinkage. While shrinkage is somewhat helpful in the absence of de-noising, it adds no benefit in combination with de-noising because shrinkage dilutes the noise at the expense of diluting some of the signal as well.

TABLE 1. RMSE FOR COMBINATIONS OF DE-NOISING AND SHRINKAGE (MAXIMUM SHARPE RATIO PORTFOLIO)

	Not De-Noised	De-Noised
Not Shrunk	9.48E-01	5.27E-02
Shrunk	2.77E-01	5.17E-02

Signal-Induced Portfolio Instability

Convex optimization solutions are unreliable for a second reason: signal-induced instability. There is an intuitive explanation for how signal makes mean–variance optimization unstable. When the correlation matrix is an identity matrix, the eigenvalue function is a horizontal line and the condition number is 1. Outside that ideal case, the condition number is affected by irregular correlation structures. In the particular case of finance, when a subset of securities exhibits greater correlation among themselves than with the rest of the investment universe, that subset forms a cluster within the correlation matrix. Clusters appear naturally, as a consequence of hierarchical relationships (e.g., the tree structure of MSCI's sector classification). When K securities form a cluster, they are more heavily exposed to a common eigenvector, which implies that the associated eigenvalue explains a greater amount of variance. But because the trace of the correlation matrix is exactly N, an eigenvalue can only increase at the expense of the other eigenvalues, resulting in a condition number greater than 1. Consequently, the higher the intra-cluster correlation, the higher the condition number. Recall that noise-induced instability results from too few observations. Signal-induced instability is caused by an entirely different reason—namely, clustering of the correlation matrix.

To address this problem, we can apply the nested clustered optimization (NCO) procedure introduced by López de Prado (forthcoming 2019): (1) Cluster the correlation matrix using an ML algorithm like the one described in López de Prado and Lewis[40]; (2) apply the optimization algorithm to each cluster separately; (3) apply the optimal weights to collapse the correlation matrix to one row (and column) per cluster; and (4) apply the optimization algorithm to the collapsed correlation matrix. The optimal weights are the dot product of the weights in (2) and the weights in (4). The reason this procedure delivers robust weights is that the source of the instability is contained within each cluster, thus rendering a correlation matrix that is well behaved.

40 M. López de Prado and M. Lewis, "Detection of False Investment Strategies Using Unsupervised Learning Methods," *Quantitative Finance* (forthcoming 2019). Available at https://ssrn.com/abstract=3167017.

We can demonstrate the efficacy of this approach through the same Monte Carlo procedure we described earlier. **Table 2** reports the results from 1,000 experiments. The NCO method computes the maximum Sharpe ratio portfolio with 45.17% of Markowitz's RMSE (i.e., a 54.83% reduction in the RMSE). The combination of shrinkage and NCO yields a 18.52% reduction in the RMSE of the maximum Sharpe ratio portfolio, which is better than with shrinkage but worse than with NCO. Once again, NCO delivers substantially lower RMSE than Markowitz's solution, and shrinkage adds no value.

TABLE 2. RMSE FOR THE MAXIMUM SHARPE RATIO PORTFOLIO

	Markowitz	NCO
Raw	7.02E-02	3.17E-02
Shrunk	6.54E-02	5.72E-02

Key Takeaways

Markowitz's portfolio optimization framework is mathematically correct, but its practical application suffers from numerical problems. In particular, financial covariance matrices exhibit high condition numbers because of noise and signal. The inverse of those covariance matrices magnifies estimation errors, which leads to unstable solutions: Changing a few rows in the observations matrix may produce entirely different allocations. Even if the allocations estimator is unbiased, the variance associated with these unstable solutions inexorably leads to large transaction costs that can erase much of the profitability of investment strategies.

De-noising and NCO ... achieve very significant reductions in the RMSE of the estimated optimal portfolios.

Factor-based correlation matrices, such as those computed by BARRA, very tangentially address noise-induced instability and do not address signal-induced instability. In this article, we have discussed two methods, de-noising and NCO, which achieve very significant reductions in the RMSE of the estimated optimal portfolios. Practitioners should become familiar with these two methods and apply them to their particular problems.

11. SPECIAL FOCUS: USING INTELLIGENT SEARCHES TO COLLECT AND PROCESS INFORMATION

Contributor: Yang Yongzhi[41]

AI engines are built using a large amount of data and the help of machine learning, computer vision, image recognition, natural language processing, and optical character recognition. They have the potential to replace much of the repetitive manual tasks that take up a large percentage of junior analysts' time.

As the internet gets bigger and bigger, it also gets denser in terms of data. Key data in current search results are either hidden between the notice or are duplicated in multiple places. AI engines are trained to understand and extract information from the toughest kind of sources, which includes social media as well as other websites, and also process locally available information.

AI engines can enable comprehensive searches of public announcements, research reports, financial news, and databases. This type of search engine goes beyond text information and uses computer vision to search graphs and tables embedded in the documents as well. It also mines for value in alternative data by developing entity relationship models by security, industry, company, individual, and media outlet.

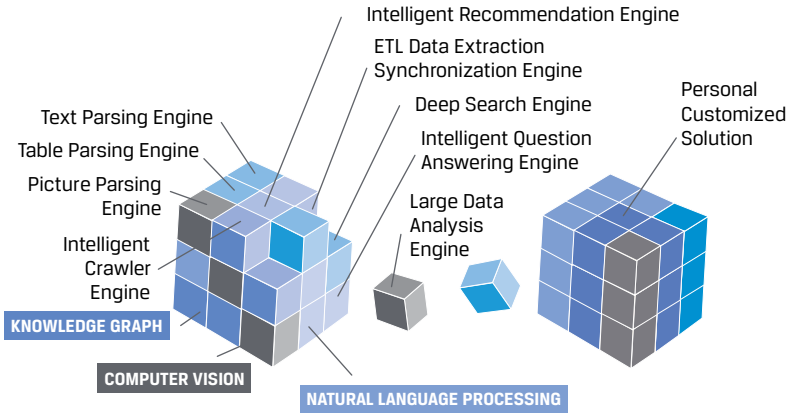
Unstructured data includes data in such formats as PDFs, images, pictures of charts and tables, social media posts, blogs, news items, and much more. AI engines allow users to search and extract key data from hundreds of sources at once. Using the power of AI, these engines also track any changes that have been made and adjust accordingly, eventually allowing analysts to automate repetitive tasks and use their time more productively.

Figure 13 provides an overview of our AI application framework.

The analysts' main output is research reports. Analysts at an average investment firm may use multiple portals to follow market developments and company information. AI engines—an example of which is illustrated in **Figure 14**—enable them to be more productive in the following areas:

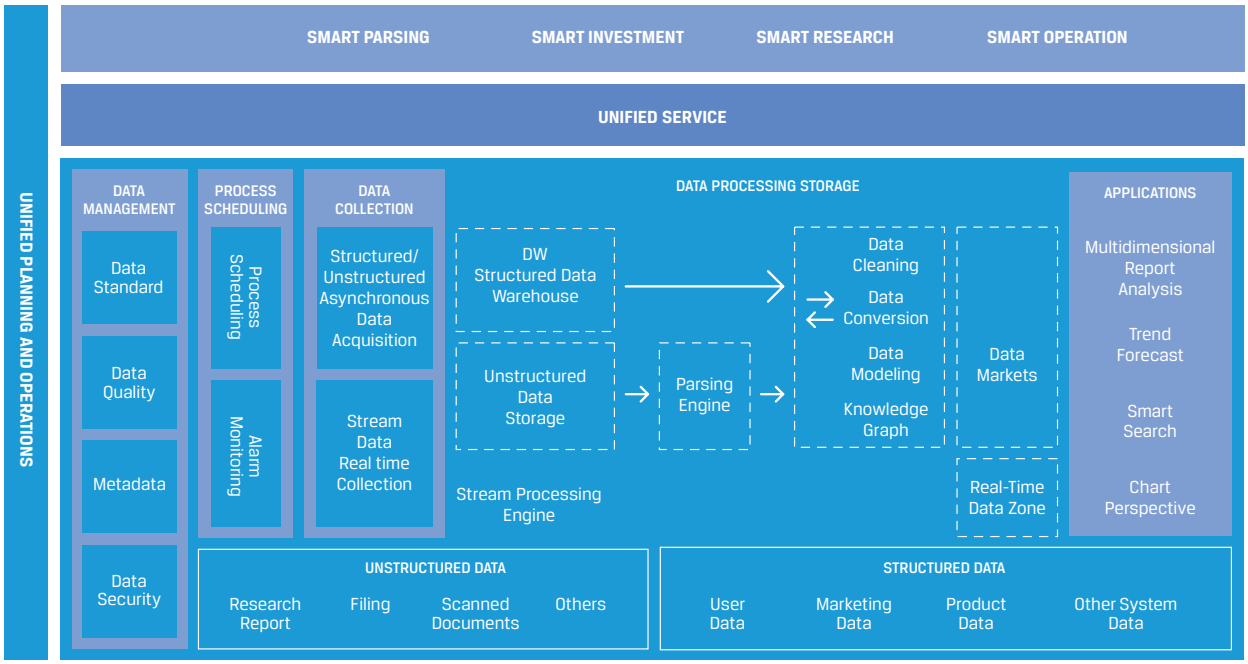
- Market developments. Analysts can now use a single platform to search and extract information from multiples sources and input the data into models in seconds.
- Alerts. The systems can usually be customized for analysts' specific research requirements, including providing alerts when new information, such as a policy change, is detected.

FIGURE 13. ABC FINTECH AI OVERVIEW



41 Founder, ABC Fintech. This case is prepared based on interviews with him and materials submitted by ABC FinTech.

FIGURE 14. ANALYST.AI FRAMEWORK



Some buy-side analysts rely on sell-side reports and public sources to produce research. AI engines can broaden the analysts' reach by allowing them to perform deeper and more comprehensive searches. For example, analysts are now able to search information embedded in reports, such as legends. AI engines also support the extraction of data from graphs and tables in reports, which has significantly reduced the time analysts spend on data gathering while improving the quality of the data collected. A report that used to take an analyst three to four weeks to produce now takes only a couple of weeks at most. Analysts are also able to spend more time conducting interviews and developing investment theses.

As an example, an analyst covering the security industry was able to detect changes in industry trends before the market by going beyond meeting with management of the incumbent companies and their competitors. The analyst saw that the focus of the industry had transitioned from the placement of security cameras to the cloud and purchasing decisions were becoming more centralized to the municipalities, counties, and provinces. The scale and complexity of the projects have increased. The trend favors new entrants with competitive advantages in technology, experience managing large-scale projects, and government relations. The analyst said he was able to see these trends because he had time to do more investigative research rather than spending most of his time collecting information.

ACKNOWLEDGEMENTS

In addition to the contributors listed under each case, we'd also like to thank the following for their helpful input in support of this project:

Atit Amin, CFA

Ernie Chan

Simon Chao

Lisa Chen, CFA

Gordon Coughlan

Jan Deahl

Rebecca Fender, CFA

Francesca Guinane

Ariel Finno

Bao Jie

Yao Jun

Hinesh Kalian

Rob Langrick, CFA, CIPM

Shu Ming, CFA

Jerry Pinto, CFA

Hugh Simon

Carol Ward

Jiang Wen

Kelly Ye, CFA

CFA INSTITUTE STAFF

Larry Cao, CFA, Senior Director, Industry Research
Author

Rhodri Preece, CFA, Senior Head, Industry Research
Editor

Gary Baker, CFA, Managing Director, EMEA and Industry Research

ISBN 978-1-942713-78-4

CFA Institute

www.cfainstitute.org
info@cfainstitute.org