HANDBOOK OF ARTIFICIAL INTELLIGENCE AND BIG DATA APPLICATIONS IN INVESTMENTS

IV. CHATBOT, KNOWLEDGE GRAPHS, AND AI INFRASTRUCTURE

This book can be found at cfainstitute.org/ai-and-big-data

9. INTELLIGENT CUSTOMER SERVICE IN FINANCE

Xu Liang, PhD
Chief Engineer, Ping An OneConnect

Development of AI-Enabled Intelligent Customer Service in the Financial Industry

The ubiquitous penetration of internet and big data applications has led to tremendous changes in consumers' behavior patterns and lifestyles, raising the expectations and requirements for financial services while steering service channels and models toward greater personalization. However, the traditional IT plus human customer service model can barely cope with the expanding user scale and service needs that have become increasingly diversified and fragmented. This situation intensifies the conflict between the limited availability of human customer services and massive volumes of customer inquiries and service needs. Financial institutions continue to add customer service positions, but the customer service function frequently falls into the operational trap of sharp increases in labor costs, fragmentation of user needs, and lower service satisfaction (Liping 2018). There is a pressing need for financial institutions to convert their customer service systems into intelligent digital systems to improve customer service responsiveness, optimize service experience, and at the same time, reduce costs and boost efficiency.

With the development of a new generation of intelligent technologies, intelligent customer service systems powered by artificial intelligence (AI) technologies, such as big data analysis, knowledge engineering, machine learning (ML), and intelligent voice, can build the bridge to interact with those using media, such as text, voice, and images, as well as to assist in human conversations, quality inspection, and business processing and, in turn, to lower the manpower costs for financial institutions and improve their service response efficiency. Compared with traditional customer service systems, AI-driven intelligent customer service offers significant advantages in various dimensions—such as channels, efficiency, and data—driving customer service centers to shift to a digital operation model powered by AI (Yuan 2021). Besides addressing the pain points of traditional customer service, AI-based intelligent customer service also helps financial institutions build closed-loop data value chains that enable the monetization of data and accelerate their transition to become smarter digital enterprises.[1]

On the one hand, the intelligent and digital transformation of customer service will significantly improve companies' operational efficiency and reduce costs to enhance cost efficiency. Intelligent customer service empowers enterprises to better identify the true needs of customers, leading to a more efficient allocation of customer service resources that improves business processing efficiency and reduces manpower costs. In addition, companies can upgrade their products and services according to customer requirements obtained from customer insights to boost a company's profitability.

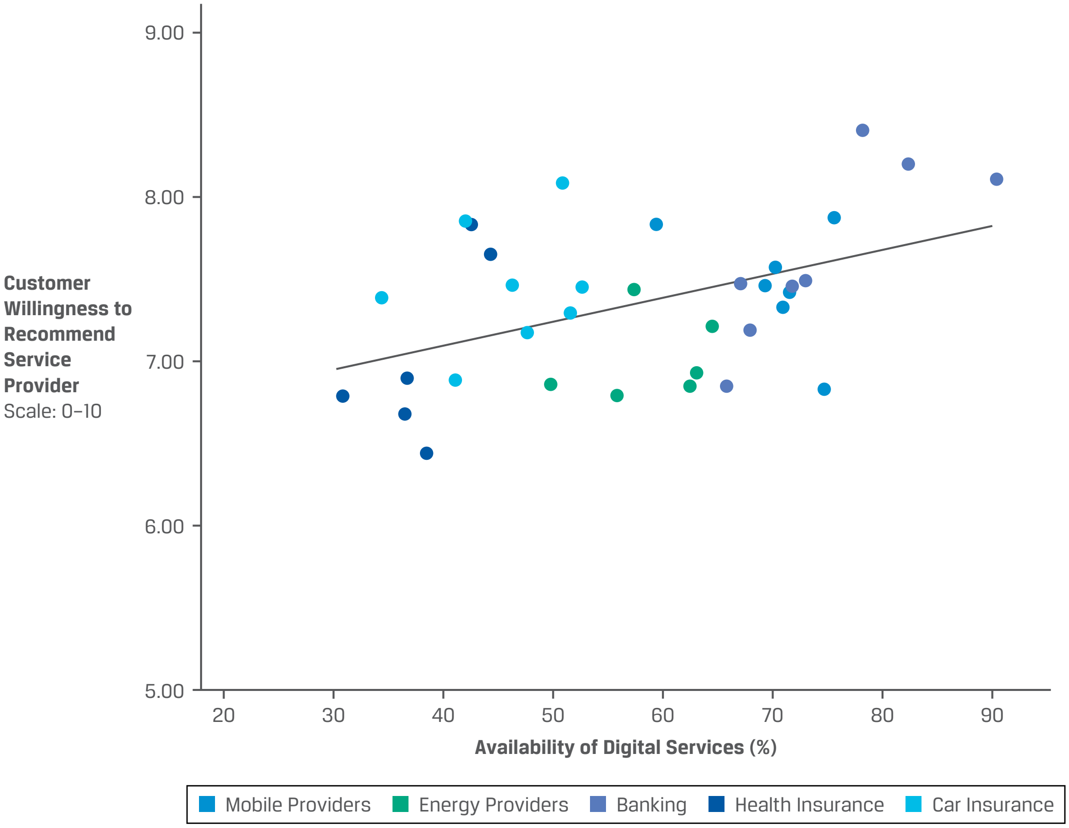
On the other hand, intelligent customer service will offer better user experience, enhance a company's reputation, increase user satisfaction, and create a differentiated brand image in its role to reinforce positive incentives in both directions. Intelligent customer service can support omni-channel services, satisfy customer inquiries anytime and anywhere, offer customers a personalized service experience, and improve customer satisfaction. Moreover, an increase in user recommendations generated from improvements in digital service capabilities will help build the brand image and strengthen the brand's competitiveness. According to a McKinsey survey conducted in 2019 (Breuer, Fanderl, Hedwig, and Meuer 2020), user recommendations of banks increased significantly with improvements in the banks' digital service capabilities. Companies in the financial industry that have improved their digital service capabilities are more likely to receive a higher level of user recommendations when compared to other industries (see **Exhibit 1**).

The COVID-19 pandemic has further spurred the transformation of financial services to a new online and contactless service mode. This creates an unprecedented development opportunity for intelligent customer service centers, and the outlook for the global intelligent customer service market is rosy. The market size of global intelligent contact centers was USD1.07 billion in 2019 and is expected to grow at a compound annual growth rate (CAGR) of 36.45% from 2023 to 2028).[2]

[1] This information came from DBS Bank's Marketplace website: www.dbs.com.sg/personal/marketplace/ (accessed 15 June 2020).

[2] For more information, see the Mordor Intelligence report titled "Intelligent Virtual Assistant (IVA) Market—Growth, Trends, COVID-19 Impact, and Forecasts (2023–2028)": www.mordorintelligence.com/industry-reports/intelligent-virtual-assistant-market?gclid=EAIaIQobChMI9Ki1i-nf_QIVuxCtBh1ZdARcEAAYASAAEgJGovD_BwE.

Exhibit 1. Relationship between Customers' Willingness to Recommend a Service Provider and Digital Service Capabilities in Various Industries



Source: Breuer et al. (2020).

The intelligent customer service industry and market in China are also booming. According to a survey by Frost & Sullivan China, the market size of China's intelligent customer service industry in 2020 was RMB3.01 billion, growing at an explosive rate of 88.1% year over year. As intelligence continues to make headway in the industry, the market size of China's intelligent customer service industry will proliferate in the next five years and is projected to reach RMB10.25 billion in 2025 (as shown in **Exhibit 2**), representing a projected CAGR of 35.8% from 2020 to 2025 (Yuan 2021).

The intelligent customer service industry is clearly experiencing a rapid growth with huge demand. Intelligent customer service systems will bring more value to financial institutions and strengthen their competitiveness. Currently, the industry is at the crossroads of divergence and reshuffling as giant players begin to emerge. Intelligent customer service is reshaping the value of the customer service center, transforming it from a cost driver to a value driver. The onset of the 5G era also brought new opportunities for the industry. Real-time interoperability, interconnection of all things, and innovative intelligent customer service application scenarios and formats combine to make real-time and intuitive interactive dialogue the goal and development direction for future R&D in intelligent customer service.

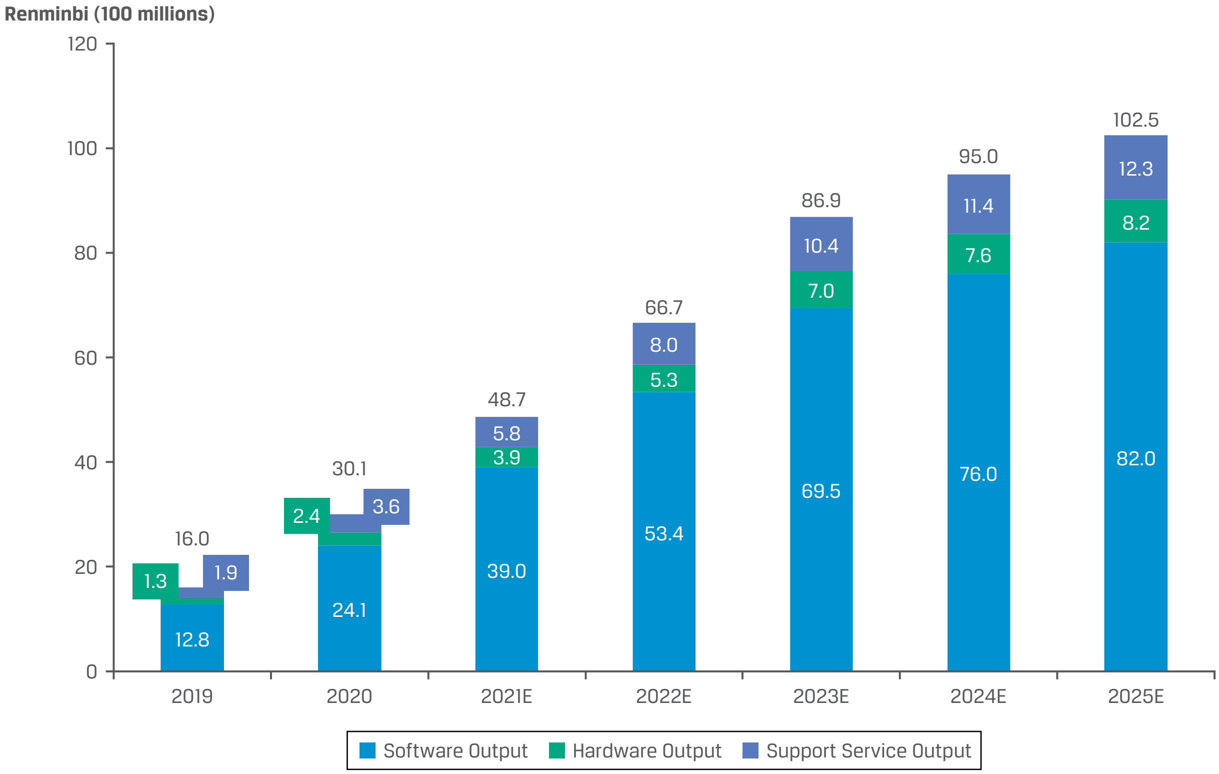
No matter where the financial institutions are in their digital transformation stage, the overall landscape of the financial industry has been fundamentally altered by the COVID-19 pandemic. In the future, financial institutions will have to compete and win in new ways.

The COVID-19 pandemic will have a far-reaching impact on future financial services in several aspects, including customer behavior, products and services, business operations (front-/middle-/back-office), operating models, external cooperation, and working methods.

- **As digital channels become the "basic offering," financial institutions must innovate to create unique customer service experiences.**

  As the COVID-19 pandemic accelerates the digitalization of front-office functions, customer experience differentiation in such areas as online account opening and login, one-click loan underwriting, and remote claim settlement will form the new fundamental requirements for customer service. Investment in

Exhibit 2. China's Intelligent Customer Service Market Size, 2019–2025E



Source: Yuan (2021).

these areas may be the key for financial institutions to attract and retain their customers. To establish a differentiated value proposition, financial institutions can further explore products and services that meet customers' non-financial needs. In the case of banks in Asia, DBS Bank's Marketplace platform allows customers to search for, buy, and sell properties and cars; book their trips; and even compare and switch their payment plans for utilities (water and electricity).[3] These trends are apparent in both retail and corporate financial services, and financial institutions have to adjust their value proposition accordingly.

Likewise, the "last mile" of digital services may be the next arena as financial institutions strive to offer more personal and human digital interactions. To this end, financial institutions need to use more data to better understand their customers in order to deliver products that meet customer needs via the right channel at the right time.

Furthermore, even as financial institutions directly assign more control to customers with digital capabilities and self-service capabilities, there is still a need to maintain the human experience with branch offices and call centers. Once viewed as a traditional solution, call centers are now becoming more important. In addition, financial institutions should speed up the transformation of their branches to achieve a balance between supply and demand. New concepts and automatic teller machines (ATMs) with more autonomy offering the suite of legacy services and functions that was previously exclusive to branches may emerge, business outlets may be further integrated, and full-service branches may be reorganized to become specialized consulting centers.

- **A focus on service costs will be key to adapting to the new industrial economy and surviving difficult times.**

The recent market contractions are likely to continue. With additional pressure on profits, digital transformation may become an important means to improve efficiency and optimize costs. According to recent surveys by the European Central Bank (ECB), the average cost-to-income ratio for digital-only banks is 47%, well below the average for smaller banks (73%). Although digital-only banks are in the early stages of development, the difference in this ratio also highlights the

[3] For more information, go to the platform's website: www.dbs.com.sg/personal/marketplace/.

many cost advantages that full digitalization can bring (Deloitte 2020).

For traditional financial institutions, bridging the technology gap aggressively will significantly improve long-term operational efficiency and accelerate industry modernization and digitization (such as public cloud and AI). Technology investments may be more challenging in the macro business and economic environment, but in the long run, the elimination of structural costs through comprehensive digital transformation will be important for traditional financial institutions as they compete with digitally savvy newcomers. To achieve this outcome, traditional financial institutions have to look beyond conventional investment horizon preferences (i.e., more than three to five years).

To sum up, despite an unknown future fraught with uncertainty, the development trend of the financial services industry that will follow the end of the COVID-19 pandemic has already been set in motion. Over the next year, financial institutions will have the unique opportunity in the retail and small business spaces to build trust and loyalty among customers by being responsive to customer needs and providing quality services.

To this end, financial institutions have to align the services they offer with the current situation and accelerate their digital and intelligent development with a sense of urgency. Most importantly, however, they need to better adapt to customer needs. Customer trust and loyalty (and even willingness to share data) will come and go depending on how well the financial institutions build trust and loyalty.

Classification and Typical Application Scenarios of AI-Based Intelligent Customer Service in the Financial Industry

With the rapid development of fintech in China, the penetration rate of intelligent customer service in the financial industry continues to rise. The key role of intelligent customer service is to assist human operators and take over repetitive basic inquiries to free up manpower. However, problems related to specialized topics often require an intervention involving both intelligent and human customer services.

Classification

Based on the specific customer service duties, intelligent customer service can be divided into three types: service oriented, investment advisory, and outbound.

Service-oriented customer service

Service-oriented customer service applies natural language understanding formed from training with a large corpus and develops context models according to specific business scenarios to conduct a natural dialogue and question-and-answer session with users. It fulfills the service requirements for a natural dialogue and addresses the customer's service needs in general scenarios, such as account top-up and transfer, as well as business issues. The service is the most common basic application scenario of intelligent customer service in the banking industry. Currently, financial institutions in China have adopted such technology in PCs and mobile terminals. On the one hand, service-oriented customer service can automatically answer many repetitive basic inquiries to relieve the work pressure of human agents to a great extent. On the other hand, it can respond quickly to customer needs as soon as the user encounters a problem, making it convenient for customers to seek solutions, which promotes service satisfaction.

As AI technology progresses, service-oriented customer service will continue to improve in terms of recognition accuracy and diversity of responses to user inquiries and even provide different answers according to different services and scenarios. For example, the Ping An Pocket Bank app can provide targeted answers to inquiries based on the type of bank card held by the customer. This capability further bolsters service quality and user satisfaction.

Investment advisory customer service

As the Chinese economy has continued to develop, disposable income per capita has gradually increased, and beyond the expenses to fulfill basic daily living needs, the demand for asset appreciation has intensified. Investment advisory customer service has emerged to meet the investment and wealth management needs of financial customers. It can use algorithms and wealth management know-how to analyze users according to different risk preferences, expectations on returns, and investment directions and match users with various products offered by financial institutions. Users get investment recommendations on the type, scale, and timing of investment products, and the system also tracks the customer's investment returns to adjust such investment recommendations accordingly. Today, more and more financial institutions are deploying investment advisory customer services. Companies can use relevant codes developed on mobile terminals to allow investors to conveniently access investment advice and analysis reports on their mobile phones. As AI technology continues to advance, there will be further improvements in the capabilities of investment advisory customer service to offer a better service experience to financial customers.

Outbound customer service

Strong demand for telemarketing in the banking industry coupled with the need to curtail rising manpower and operating costs translate to increasing demand for outbound customer service. The conventional outbound customer service has to deal with not only manpower cost constraints but also extensive training for the agents, which drains a lot of resources. Moreover, data have demonstrated that the success rate of traditional telemarketing is relatively low. With technology advancements, the banking industry has initiated many studies on intelligent outbound customer service. The service can verify the user identity and, at the same time, provide users with simple feedback and guidance to complete the relevant sales and marketing process. Continuous technology optimization enables outbound customer service to gradually strengthen language skills, allowing it to offer the same dialogue capability and service experience as human agents and handle tasks that were originally performed by humans to save the related costs.

Typical Use Cases

Intelligent customer service is mainly applied in two scenarios: presales services and after-sales services. Presales services are mainly deployed in telemarketing scenarios and used to improve call efficiency, data security, and service quality control. After-sales services are mainly for customer consultation and service callbacks, and the key is to meet the accuracy requirements of outbound products.

By sector

Different sectors in the financial industry, such as insurance, banking, internet finance, and securities, have varying requirements for intelligent customer service due to their inherent business characteristics (see **Exhibit 3**).

Insurance

Insurance is a relatively complex business; every user has unique requirements from insurance purchase to claims. Therefore, the insurance industry has relatively high requirements on the intelligence level of the intelligent customer service, which has a direct impact on customer satisfaction, retention, and conversion.

Banking

The main banking applications are in handling highly repetitive queries, providing response services during peak periods, and providing intelligent collection services. Intelligent customer service mainly assists human customer service agents and frees up manpower during peak periods.

Fintech

Similar to banks, customer service in the fintech industry has to deal with a large volume of repetitive inquires, so intelligent customer service is mainly deployed to handle highly repetitive queries, with the intelligence focused on such areas as intelligent debt collection and smart marketing.

Securities

As with the banking industry, customer service in the securities industry also has to deal with a large volume of repetitive basic inquiries. Here, intelligent customer service is mainly used to address inquiries on basic problems in the securities sector, with the intelligence applied in smart stock selection and analysis, as well as smart push services.

Exhibit 3. Application and Business Scenarios of Intelligent Customer Service in Various Financial Sectors

Insurance

Enquiries on insurance-related issues such as insurance purchase are relatively complex, and the level of intelligent service will affect customer retention and conversion.

Banking

The main applications are in handling highly repetitive queries, providing response services during peak periods, and supplying intelligent collection services.

Internet Finance

Similar to banking, the main application is in handling highly repetitive inquiries. Intelligence is focused on intelligent collection and smart marketing.

Securities

It is mainly used to address inquiries on basic problems in the securities sector, with the intelligence applied in smart stock selection and analysis, as well as smart push services.

Within sectors

Within sectors (in banking, for example), the application scenarios of AI can be divided into the following six categories.

Robots in intelligent voice navigation and service scenarios

Application concept: To the extent that customers are receptive to robots, voice robots are used in navigation and service scenarios to improve service quality, reduce costs, and improve efficiency.

Today, applications using intelligent voice navigation robots have matured and are widely adopted by large banks. However, it is still possible to further expand the application of this technology. In terms of intelligent voice applications, due consideration should be taken to incorporate customer complaints, customer acceptance, and scene adaptability. Aggregate analysis based on the two dimensions of customer attributes and scene responsibility should be carried out to identify the most suitable business scenarios for robot applications and the type of customers who are most receptive to robots. Robots can be gradually introduced to eventually encompass all scenarios to reduce costs and improve efficiency. At the same time, customers' acceptance of robots has to be examined for service quality.

Smart identity verification with voiceprint

Application concept: Reduce costs and mitigate risks by using voiceprint to identify customers to reduce verification time, shorten the overall call duration, and retrieve historical recordings to identify counterfeit and fraudulent voiceprints.

Currently, the application scenarios and technical practice are still being developed for this technology. The recommended approach is to supplement the identity verification process with voiceprint technology when the service is being performed. With the permission of the customer, a customer voiceprint record is established that will be used together with other identification techniques to seamlessly verify the customer's identity for any incoming service request or inquiry from the customer. This will promptly identify incoming calls from parties impersonating the customer while reducing the time taken to verify the customer's identity to improve the customer experience.

Intelligent quality inspection

Application concept: Improve cost efficiency by substituting labor with quality inspection robots to perform objective and comprehensive inspections, reduce the need for manual and repeated playing of recordings, and pinpoint problems quickly.

Currently, this application technology and scenario have already matured, but there is still room for enhanced technology empowerment. An offline quality inspection system is suitable for service optimization. Once a customer problem is found, the offline quality inspection model can quickly locate similar problems in the database and determine the appropriate improvements. Real-time quality inspection is suitable for correcting customer service habits. It monitors the dialogue process between the customer service agents and customers, identifies the problems, and provides feedback promptly to correct mistakes by the customer service agents.

Intelligent human–machine collaboration

Application concept: Reduce costs and improve efficiency by analyzing the interaction linkages of the scenarios in the process flow and select business scenarios that can be substituted by robots.

The specifications for such applications are currently being developed. To engage customers using a mix of robot outbound call and human collaboration, start from the business operation process, subdivide the key interaction links of all scenarios in the end-to-end process, as well as the characteristics of customers who called, and match with the robot capabilities one by one to select the business scenarios and customer types that the robots can serve to improve customer satisfaction and business conversion rates.

Online service and sales

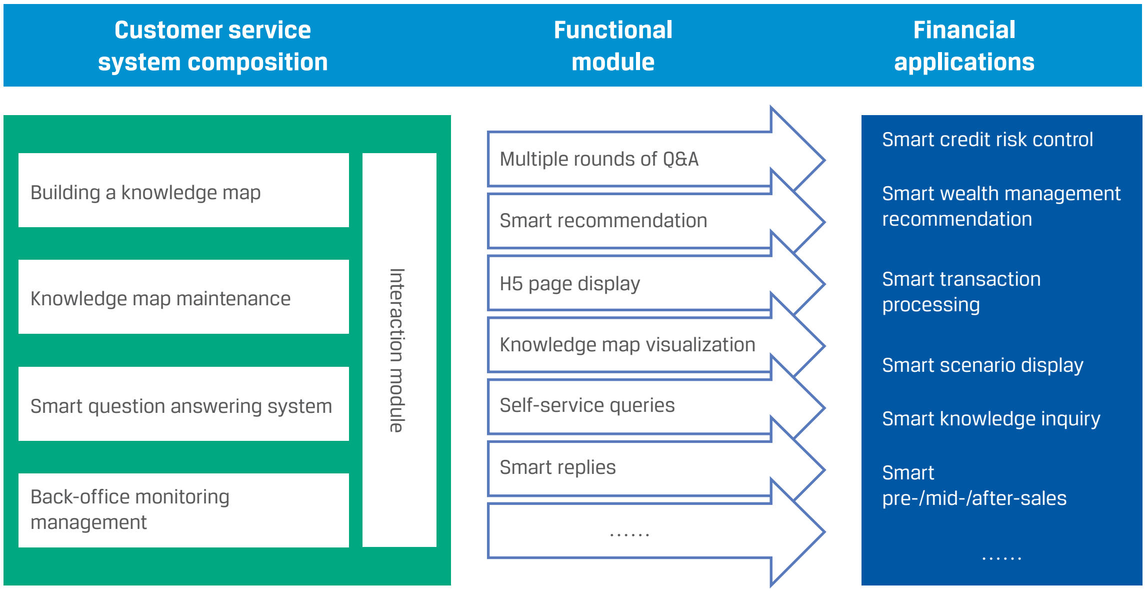
Application concept: Based on the text robot or chatbot, online service and sales use big data analysis and product configuration to recommend interactions for the center. It provides precise recommendations on touch points to complete service and offers sales and operations with better performance and precision—and thus manages to improve user experience and contributes to business indicators.

Customer centers and remote banks have higher requirements for text robots. The systems should be designed for flexibility to offer more diversity in service formats and facilitate operations and maintenance for greater convenience and speed. The goal is to achieve rapid coverage of service channels (such as WeChat, Weibo, and TikTok) and build capabilities to integrate services with marketing efforts.

Smart knowledge management

Application concept: Establish a knowledge management platform that supports all channels (agents, robots, customers), forming a centralized maintenance system to reduce repeated purchases and investments as well as to cut costs and improve efficiency.

Exhibit 4. Intelligent Customer Service Architecture for the Financial Domain


Customer service system composition
Functional module
Financial applications
Building a knowledge map
Knowledge map maintenance
Smart question answering system
Back-office monitoring management
Interaction module
Multiple rounds of Q&A
Smart recommendation
H5 page display
Knowledge map visualization
Self-service queries
Smart replies
......
Smart credit risk control
Smart wealth management recommendation
Smart transaction processing
Smart scenario display
Smart knowledge inquiry
Smart pre-/mid-/after-sales
......

The application should standardize the method used to organize the knowledge based on the service characteristics and scope, guiding each department to break down the knowledge systematically to implement the fragmentation, structuring, and labeling of scenario-based knowledge content.

Building a Financial Intelligent Customer Service Process

The development of financial intelligent customer service often requires the active participation of business analysts and implementation personnel in the early stage, online optimization of the customer service system, in-depth access to the businesses, and the organization of the business processes and user inquiry scenarios for a tight integration of the business and intelligent customer service functions with customizable configuration for the processes. The development of a financial intelligent customer service system can be divided into two parts: system construction and operation optimization.

System Construction

Before a financial institution builds an intelligent customer service system, first it needs to define the architecture of such a system by organizing the structure of the system and clarifying the composition, functional modules, and business applications that must be supported for the system. After which, it can carry out the work to implement the customer service system by technological means.

The intelligent customer service architecture for the financial domain is shown in **Exhibit 4**.

Currently, the intelligent customer service architecture for the Financial Domain consists of at least five components: interactions, graph construction, knowledge graph maintenance, intelligent Q&A, and back-office management. More functional modules can be derived from the interactions module, such as multiple rounds of Q&A, smart recommendation, H5 page display, knowledge graph visualization, self-service queries, and intelligent replies. Corresponding intelligent financial applications can be derived from a combination of various functions and embedded interactive formats, and these include applications for credit risk control, financial management advice, business processing, scenario display, knowledge search, presales, sales, and postsales.

The deployment modes for intelligent customer service include public cloud, hybrid cloud, and private deployment. The main differences between the three modes are reflected in such areas as data privacy, construction cost, and deployment cycles. Different modes meet the needs of different customers.

Of the three modes, deployment using the public cloud has the lowest construction cost and shortest deployment cycle, which is suitable for SMEs (small and midsize enterprises) that want to quickly build their intelligent customer service capabilities. On-premise deployment requires enterprises to establish their own technical teams; although the construction cost is high, such a deployment greatly enhances data privacy and security, which are suitable for large enterprises. A hybrid cloud deployment combines the characteristics of the previous two deployment methods. Although it uses the public cloud platform, the data are stored locally, and such a deployment may become the mainstream deployment mode for intelligent customer service in the future. Currently, on-premises deployment is still the market mainstream, but cloud deployment is increasingly favored by the market owing to its low cost, fast deployment, and agile iteration.

An excellent intelligent customer service system should have expertise in knowledge graph as its basis, supplemented by advanced speech semantic recognition and understanding technologies to deliver an outstanding interactive experience. At the same time, it can fulfill as many business functions as possible and enable access to the capabilities of many platforms and many channels so as to quickly deploy customer service capabilities in different business environments and support a variety of financial scenarios.

Operational Optimization

An intelligent customer service system is not built overnight. Many financial institutions are not able to achieve the business results that their intelligent customer service systems were projected to deliver after the systems were built or purchased. A major reason is that financial institutions' operational processes are generally not designed with intelligent customer service in mind and are not able to optimize and improve their services. A good intelligent customer service system is often the result of cooperation between products and business personnel.

The intelligent customer service product has to be supported by cumulative business expertise and experience in AI technology. Business operators need to listen to service voice recordings to understand recent user demands and be familiar with the business. They also need to be familiar with various channels, have an acute business sense, maintain communication with business departments, and understand the current business, the drivers behind the business, and the development trend of the business. They also need to acquire and constantly update their business knowledge and adjust the corpus in time to ensure that the customer service products can meet evolving business needs. At the same time, operators need to leverage business data to uncover existing problems in products or new demands in the industry and constantly iterate the intelligence level of the intelligent customer service system to improve the service capability to address inquiries.

A good vendor for intelligent customer service products, besides having advanced technologies, should also have extensive industry experience and a standard training and certification system to help financial institutions build an intelligent customer service system that consists of not just the hardware but also the soft power that can be derived from operating the intelligent customer service product, helping financial institutions gain a true understanding of AI so that they can leverage AI in intelligent customer service to achieve the twin goals of cost reduction and higher efficiency.

Financial AI Customer Service Case Studies

Financial institutions' diverse needs in serving their customers are increasingly met by a slew of AI-integrated applications merging voice products with AI technology.

Inbound Call Middle Office

Companies face many challenges in building their own AI products, such as repetitive development efforts, the pace of adoption not keeping up with business needs, and poor robot performance. Inbound call middle-office solutions can help enterprises build intelligent customer service systems faster and better by addressing these issues.

The middle-office solutions have proven effective in training new customer service agents and adapting existing AI solutions for new business. On the back end, these middle-office solutions can connect with and manage AI engines, such as natural language understanding (NLU) and voice recognition robots, from many vendors. Similarly, on the front end, they can accommodate inbound inquiries from multiple channels, such as phones or PCs. They also offer extensive business solution modules and allow clients to quickly build on their existing capabilities and adapt to changing business needs.

In one case, OneConnect's Gamma Mid-Office Voice Access helped a large Chinese bank simultaneously launch an intelligent assistant and an intelligent tutor within one month. The solution adopted the same underlying engine to offer two products with a one-time development effort. By resetting a variety of robot rules, the solution enabled the robots to fulfill new business scenario requirements, distinguish between scenarios, and accommodate the needs for both scenarios.

Intelligent Outbound Call

Traditional customer service has drawbacks, such as substantial manpower investment, long training duration, poor work efficiency, performance improvement challenges, and complex data analysis. There are three major customer service pain points in the industry. The first is business reminders. Financial institutions have a need for a lot of information synchronization and business reminders, such as credit card payment reminders and reminders about dues for financial products. Sending manual reminders will occupy 40% of the daily working hours of an operator and usually involves repetitive tasks that lead to low work enthusiasm and worsening customer experience. The second pain point is financial collection. Take the credit industry as an example. Every year, about 70% of overdue accounts could be attributed to account holders who forgot their repayments, and most of these amounts could be recovered after manual reminders and notifications are sent. However, manual collection is time consuming, with low work efficiency and poor time utilization. The third is active marketing, where sales personnel often have to filter through a massive number of leads for financial products, especially insurance products, which is a time-consuming and labor-intensive task that tends to undermine sales confidence. The inconsistent grading of customer intentions and the loss of high-quality customer leads are two problems that result in the low efficiency and high cost of active marketing.

The pain points in the previous scenarios call for an intelligent outbound call solution. With the call center system as the base, one such solution layers on a number of AI technologies, such as natural language processing (NLP), speech recognition, and speech synthesis, replacing manual calls with intelligent outbound robots to screen intended customers, target customers, and accurately classify different customers. By effectively reducing headcount, the product helps companies improve the customer experience and increase their marketing efficiency as it steers communications with customers toward greater professionalism, automation, and efficiency. In actual operations, the results are promising.

The specific process of an outbound robot is as follows:

1. First, the customers may call customer service, which is available 24/7. The efficiency is high, service availability is long, and incoming customer inquiries can be attended to at anytime and anywhere.
2. The customer service function will then respond intelligently, and supported by automated speech recognition (ASR), NLP, and text-to-speech (TTS) technologies, the robot can understand and respond to inquiries. In short, the intelligent customer service can identify the user's question, match the corresponding answer from the database, and provide the answer automatically.
3. The intelligent customer service then automatically classifies users according to the contents of the call, helping companies accurately target customers and acquire a better understanding of the customer's inquiry and pain points.
4. In addition, during this process, customer service will record the dialogue and store the complete recording and dialogue text, making it convenient to view and track the conversation. Furthermore, the complete dialogue recording is collected and trained by AI to further iterate and upgrade the AI products.
5. Customer service will conduct additional data analysis and clearly present several data indicators to help analyze the call performance and customers' behavior.
6. Lastly, customer service will automatically optimize itself, apply deep learning, and continuously add to the corpus to improve its Q&A performance.

In marketing, the system leverages big data to generate a 360-degree customer profile, locate target customer groups, and predict the customer's propensity with respect to loans, wealth management, and insurance products. When appropriate, the system can initiate conversations.

In risk management, the system collects user data for anti-fraud surveillance and to assess borrowers' willingness and ability to repay before the due date. Voice data are also collected to predict the repayment risk.

In investment advisory, the system automatically answers questions about loans and insurance, intelligently analyzes customer needs, assists in drafting optimal investment plans, and addresses after-sales questions.

For notifications, it provides timely payment reminders, personalized communication playbooks, multidimensional configuration logic trees, and intelligent external collection.

In internal and external customer service scenarios, the applications include external marketing to customers, collection, credit checks, internal employee inquiry hotlines, and outbound calls for surveys.

Such systems have proven to be highly valuable in production environments. As an example, a system developed by Ping An OneConnect significantly improved the response rate when used in promoting a new loan product that is part of Ping An's inclusive financial service offerings. Prior to the robot deployment, the outbound call capacity of the customer service center was about 150 calls per person per day, which did not meet customer demand during peak marketing periods. In terms of performance, the inquiry rate was only about 1% for customers in the first follow-up and about 5% by the fifth follow-up. The operating cost of the marketing outbound call center remained high, at RMB150 per customer after apportionment. After the robot was deployed, the call center achieved an average daily outbound call capacity of 1 million, with a 5% marketing response rate, while the outbound marketing cost for the same batch of customers was reduced to under RMB1 per customer. When used for loan collection, the robot can make about 900–1,000 calls a day to owners of overdue accounts, and it increased the repayment rate by 30% within a week.

Text Robots

In the past, AI customer service suffered from business pain points, such as low Q&A accuracy, support for only a single round of Q&A, inability to conduct business processing, and high knowledge base maintenance costs. The latest solutions use a more advanced NLP algorithm to gain deep semantic understanding and perform semantic analysis of the context, enabling robots to conduct multiple rounds of dialogue with users. These solutions may also adopt unsupervised ML and supervised ML to minimize maintenance costs and keep the knowledge base up to date.

Business process

- **Cloud/on-premises deployment:** The cloud version allows for quick access at a reasonable cost, allowing users to rapidly deploy intelligent Q&A, whereas the on-premises deployment offers better security and business privacy for the business environment.
- **Knowledge configuration/model training:** Data modules such as FAQ, dynamic menu, simple tasks, hot topics, and welcome messages are configured in the knowledge base, offering a set of rich content with diverse functions, while Q&A accuracy is validated by model training and running tests.
- **System use:** The system is used in intelligent robot Q&A, provision of various data reports for post-loan management, and optimization of the model for education quality inspection and marking.

Business functions

- **Powerful Q&A capabilities:** The text robot system can answer questions accurately, achieve accurate contextual understanding of speech, and easily handle complex interactive questions; in addition, one Q&A portal can support seamless switching between multiple knowledge bases. Human–machine cooperation can also be implemented to support two-way switching between robots and human agents.
- **Convenient back-office management:** The system will systematically carry out smart knowledge management, regularly perform quick model training and robot education for AI robot customer service, and support personalized configuration to satisfy the needs of different customers.
- **Efficient operation outcome:** Compared with human agents, advanced customer service solutions do not require training and can be launched quickly at an affordable cost. Moreover, the robot solutions can provide services around the clock and handle many customers at the same time. In terms of compliance, the text robot system is 100% controllable, stable, and standardized, requires no emotional care, and will not get tired of answering questions. At the same time, it can record full logs and various report data for business analysis.

Business characteristics

- **Full coverage of customer service marketing scenarios:** WeChat, official account, Weibo, H5, web, and other omni-channel coverage; Q&A scenarios involving presales, current selling processes, and after sales. The PaaS platform design facilitates business system connection, offering 24/7 high-quality service.
- **Leverage on Ping An Group's years of industry experience:** The system has a comprehensive knowledge base that includes knowledge about finance, banks, and smart cities, which empowers it to provide data support for algorithms and to quantify operational indicators.
- **Advanced technology:** 100% owned intellectual property rights with an average accuracy of 85% that offers outstanding performance and high availability and is scalable and secure.

Application scenario

The Xiaoyi text robot can answer such questions as service inquiries on how to proceed to complete the process, outlet information, product function introduction, and fees. For business processing, it supports credit card application, FX transaction, loan prepayment, and card replacement, among others. As for sales recommendation, it can offer product and fund recommendations for wealth management, as well as assist in loan processing, preliminary screening for loans, and other services.

The Xiaoyi text robot has been deployed in many banks, where it mainly handles three types of service: business inquiry, business process, and sales recommendations. To date, it has delivered excellent results. For example, the Ping An Bank app has an average monthly traffic of 13.2 million (customer inquiries), with a question recognition rate of up to 99.47%. The Lufax app has an average monthly traffic of 500,000, with a question recognition rate of up to 97.95%.

Risks and Challenges of Financial AI Intelligent Customer Service Application

In this section, I will discuss the risks and challenges involved in applying intelligent customer service in the financial industry.

Risks

AI intelligent customer service in the financial industry will facilitate the offering of products and services and save manpower costs. But owing to the constraints imposed by the current development state of related technologies, there are many potential risks in the intelligent customer service business.

Business risk from misinterpretation of intent

Intelligent customer service needs to be able to identify and answer the questions raised by customers. However, because the technology for NLU is still in its infancy, the AI solution may not fully understand the questions asked by customers the way the human agents do. Once the intention is misinterpreted, the system may not find a match in the knowledge base or may match a wrong answer. There is the risk of not being able to answer or providing the wrong answer and operating instructions, which may lead to customer dissatisfaction or even losses suffered by the customers, creating reputation risk for banks. As such, the usual solution is to deal only with questions recorded in the intelligent customer service database and hand over questions that cannot be identified or that are beyond the scope of the database to human agents. This solution reduces manpower wastage and improves the user experience.

Risk of declining customer satisfaction

Compared with AI intelligent customer service, human customer service can more accurately understand customer intentions, grasp customer emotions, and communicate with customers more patiently and humanely. Based on the current statistics on human customer service, the service satisfaction level for the majority of human customer service is above 90%. Constrained by inadequacies in current intelligence technology, in practice, the intelligent customer service solution is unable to provide services like an actual person can and cannot grasp the feelings of customers in time. When customers are forced to use intelligent customer service to relieve manpower pressure, it is likely to worsen the customer experience, increase the risk of declining customer satisfaction, and even lower customer stickiness, leading to losses for financial institutions. As such, financial institutions are taking a deep dive into applications involving deep learning networks so that intelligent customer service solutions can learn to handle customers' emotions to give users a more humanized experience. At the same time, financial institutions continue to hand over customers with emotional issues that cannot be handled by AI to human agents to ensure a more satisfactory user experience and to reduce manpower cost.

Business risks associated with outbound customer service

Although outbound customer service will improve operational efficiency and reduce costs, because the technology is immature, outbound customer service currently supports only simple interactions with customers in most cases, which may lead to misunderstandings by customers. Owing to the unique nature of banking products, improper understanding may cause the client to lose assets and lead to unnecessary disputes between the client and the bank. As such, the intelligent outbound customer service also faces database limitations. If the user's question is not found in the AI database or cannot be identified by AI, the question will be automatically transferred to a human agent.

Risks from outsourcing of intelligent customer service products

AI technology is considered a cutting-edge technology that is developing rapidly, and the banking industry has not fully mastered its core content. Therefore, most customer service robot projects have to be outsourced to third-party institutions, which increases the dependency on third parties and uncontrollability for the banking industry. The bank relies on the source code of the outsourcing vendor for the transformation or project improvement, which weakens the bank's controllability. Meanwhile, because a large volume of data has to be provided to the third-party institutions for model training and user privacy is often embedded in such data, any disclosure may cause substantial risks.

Challenges

AI intelligent customer service is a comprehensive product solution that integrates industry knowledge, databases, speech recognition, NLP, and other technologies. Since there is still much room for improvement in intelligent technologies, such as NLP, and the sensitivity of financial industry data makes the industry knowledge graph less complete, there are challenges in deploying AI intelligent customer service in the financial industry.

AI technology is inadequate

A breakthrough is needed for better capability in interpreting intentions. AI customer service has low answer accuracy and poor customer satisfaction levels, supports only a single round of Q&A, and is not sufficiently responsive to questions raised in a continuous dialogue. It has simple functions, supports only dialogue inquiry, and is unable to handle business processes. In short, intelligent customer service answers lack a "human touch." As such, financial institutions should focus on the application of cutting-edge AI technology and ensure the rapid iterations of AI technology to satisfy the more demanding service requirements of users.

Stringent financial data security requirements and incomplete knowledge graph

The high costs of deployment and knowledge base maintenance lead to a relatively low cost performance presently. However, with further iterations and development of AI algorithms and the wider adoption of AI applications, the cost of intelligent customer service will gradually decrease, which will bring about better results.

Improving industry standards

With the gradual adoption of intelligent customer service, some problems have begun to surface. Intelligent customer service's ability to operate around the clock has created various social problems, such as harassment and privacy infringement, which are not conducive to the healthy development of the industry. As such, AI customer service leaders should jointly develop a single standard to ensure that all AI customer service products can fulfill the minimum industry requirements.

Development, Innovation, and Prospects of the Financial AI Intelligent Customer Service Industry

In this section, I elaborate on various aspects of the future of intelligent customer service in the financial industry and discuss its development, innovation, and prospects.

Development and Innovation

Financial AI intelligent customer service systems will continue to innovate in three aspects: technology, interaction, and business.

Technological innovation

The financial knowledge graph is in the building stage, and the potential linkages between knowledge units can be uncovered using such algorithms as knowledge reasoning, which can then be used to form a basis for potential customers' acquisition, new financial services development, and financial services promotion. AI technology is expected to continue to iterate rapidly, creating smarter and more capable knowledge applications and response mechanisms.

Interactive innovation

With the rapid development of rich media, the modes of information dissemination have become more diversified. Most of the knowledge in the financial industry is described in words or numbers; the format is simple, but with poor interpretability. By leveraging the intelligent customer service system in the rich media era, customer service products can enable various ways in which answers are presented, such as animations and short videos, to demonstrate the correct answers to users more clearly and intuitively. It can build on the advantages of knowledge graph to enhance the fluency and compatibility of dialogue, improve the system's ability to identify customers' feelings and real intentions, and quickly gear up the capability to support multiple rounds of Q&A. Innovations in interaction will further elevate users' perception of the company and increase user stickiness.

Business innovation

With intelligent customer service, financial institutions can change the current practice of independently selling a single product. In the future, with further breakthroughs in such technologies as AI and big data, intelligent customer service will be able to grasp customer intentions more accurately and provide customers with more personalized business assistance and product portfolios, shifting the business model from a product-oriented approach to a user-oriented approach. In addition, AI customer service will be deployed in various businesses of the financial industry. It will adapt to the changing business characteristics and be able to deal with the pain points in the user's current business for efficient revenue generation.

Development Trends

There are three trends in the application of AI technology in banking customer service centers and remote banking: popularization, personalization, and diversification.

Popularization

As AI applications mature, their popularity will gradually spread to the whole industry and all fields. With escalating manpower costs, the staffing cost of call centers and remote banks will also increase over the years. Cost reduction and efficiency improvements have always been major goals of each call center and remote bank. By making use of applications based on AI, such as text robots, intelligent voice navigation, and intelligent quality inspection, customer service centers and remote banks can gain a major arsenal that they can adopt to reduce costs and boost efficiency.

Personalization

As users continue to fine-tune their needs, the increasingly demanding user requirements in terms of experience are driving the transformation of customer service centers and remote banking solutions toward greater personalization. With intelligent voiceprint recognition, online service/sales, human–machine collaboration, and other applications, we can gradually upgrade from the traditional human agent model to an intelligent customer service model for comprehensive, efficient, and intelligent operations and services, which will be key to user satisfaction.

Diversification

Relying on strong technical support and market demand, the interaction between traditional customer service centers and customers will transform from straight-line transactions to multichannel and multidimensional dialogue and exchanges. Various ways of communicating appear, including videos, gestures, and even virtual reality, in addition to traditional text and voice. Such communication methods are more natural and may be a real breakthrough beyond the existing service model, such as phone calls to provide customers with faster three-dimensional services using interactions from all media, bringing about a more diversified experience and creating more distinctively emotive services.

Future Prospects

In the era of AI and big data, an increasing number of financial institutions have begun to develop financial technology. As operating costs continue to increase, it becomes even more pressing to reduce manpower costs. Yet at the same time, the demand for personalized services for customers is gradually expanding. All these issues prompt financial institutions to have additional requirements for intelligent customer service. Looking ahead, AI intelligent customer service will become even more common in the financial industry. Upgrading of talents and technologies, continuous enrichment of industry applications, and continuous improvement of industry specifications will promote the healthy development of the intelligent customer service industry.

Strengthen talent pool and technology R&D

The intelligent customer service system relies on emerging and cutting-edge AI technologies, such as speech recognition and NLP. To better optimize and develop the technologies, businesses need to continuously strengthen the talent pool, adopt best practices at home and abroad, and groom talents with financial backgrounds and information technology expertise. This is the only way to achieve a higher level of intelligence, improve the accuracy of answers to customer inquiries, improve user experience, and increase satisfaction. Going forward, with support from neural networks and big data, such technologies as knowledge computing, multimodal fusion, and privacy computing are expected to form the basis for intelligent analysis and decision making at financial institutions and to make autonomous learning of financial market conditions a reality. Such systems will also be capable in trend projections and can even provide customers with intelligent suggestions for their decision making. An efficient and accurate trading system will provide the technological power to drive innovation and development in the financial industry.

Expand the applications of the intelligent customer service system

There is a need to focus on current technology R&D and promote comprehensive and all-around coverage of application scenarios by optimizing the basic functions for a large-scale intelligent customer service platform geared mainly to mobile terminals. In addition, the interactive mode can gradually evolve from traditional voice and text to video, augmented reality and holographic projection, and other formats by combining emerging media with traditional media.

Meanwhile, AI customer service should be deeply entrenched at several levels. In marketing, AI customer service is included to generate intelligent marketing scenario applications and personalized services for individual customers that can greatly improve the sales success rate for wealth management products and promote the development of inclusive financial services. In customer service, AI customer service will tap into more multiscenario applications, such as identity recognition, intelligent customer service, and intelligent claims settlement, to serve more customers during peak hours and fulfill demand in an effective and accurate manner that can improve customer satisfaction significantly.

Develop and refine relevant industry specifications

For this emerging industry system, regulatory authorities need to keep up with trends, draft corresponding regulatory plans and regulatory processes, avoid and prevent series of risks that emerging intelligent customer service systems may bring, promote the intelligent customer service systems, and optimize the development of smart finance. Building an intelligent customer service ecosystem with multiple parties will make it an important focal point in banking and financial services.

Conclusion

In view of the COVID-19 pandemic and the digital transformation wave, financial institutions need to reshape their business models and adapt their working methods to meet ever-changing customer and market demands quickly and nimbly. Looking ahead, financial institutions must take advantage of the cultural and behavioral changes brought about by the pandemic to accelerate their digital transformation process, even as they remain cautious to avoid overreliance on existing models once the pandemic is over. This move is critical for financial institutions to weather the storm and thrive in the emerging new normal.

Riding on the COVID-19 pandemic as a catalyst for digital transformation, financial institutions can also take this opportunity to reflect on their role in emerging ecosystems and how they can create value.

The intelligent customer service system enables the online transfer of human customer service workload by integrating current AI technology, NLU, and other emerging technologies. At the same time, with point-to-point communication with customers, it can improve customer experience and efficiency significantly. However, it is also necessary to fully understand the imperfections and dual characteristics of technology and to not blindly follow the trends in online services and intelligence. In certain banking scenarios, for example, intelligent customer service is still not able to replace human agents. It is important to recognize the possible risks of the intelligent customer service system to avoid the risks and plan rationally. However, there is an opportunity to leverage the power of technology to create a new model for the development of the banking industry that is more efficient, accurate, and convenient.

Within the supervision framework of the authorities, in-depth application of intelligent customer service can be used to develop innovative financial products, change business methods, and optimize business processes to ensure the safety of customers' assets, optimize financial investment experience, and reduce investment risks in this rapid transition to a modern digital financial system that is highly adaptable, competitive, and inclusive.

References

Breuer, Ralph, Harald Fanderl, Markus Hedwig, and Marcel Meuer. 2020. "Service Industries Can Fuel Growth by Making Digital Customer Experiences a Priority." McKinsey Digital (30 April). www.mckinsey.com/capabilities/mckinsey-digital/our-insights/service-industries-can-fuel-growth-by-making-digital-customer-experiences-a-priority.

Deloitte. 2020. "Realizing the Digital Promise: COVID-19 Catalyzes and Accelerates Transformation in Financial Services." www.deloitte.com/content/dam/assets-shared/legacy/docs/gx-fsi-realizing-the-digital-promise-covid-19-catalyzes-and-accelerates-transformation.pdf.

Liping, Xu. 2018. "AI-Driven Intelligent Customer Service." Shanghai Informatization.

Yuan, Xucong. 2021. "2021 China Intelligent Customer Service Industry Insight." LeadLeo Research Institute.

10. ACCELERATED AI AND USE CASES IN INVESTMENT MANAGEMENT

Jochen Papenbrock, Doctorate

Head of Financial Technology EMEA, NVIDIA

Introduction

A growing number of investment professionals are building up capabilities and resources to exploit artificial intelligence (AI) and big data in their investment processes systematically, using the most advanced technologies and infrastructures, like "factories" for AI and simulation. This trend results from the realization by many that AI and big data will give them access to diversified sources of alpha, more effective risk management, better client access, and customization opportunities. Embracing such technologies has become a clear differentiator.

In this chapter, I will take you behind the scenes, provide insight into the art of the possible, and discuss how to best make use of the most recent technologies, such as accelerated computing platforms in an implementation roadmap. I will also provide useful tools and describe their implementation in several real-world use cases, such as diversified portfolio construction and environmental, social, and governance (ESG) investing, using such techniques as natural language processing (NLP), Explainable AI (XAI), and Geospatial AI.

Learning how to implement AI technologies is highly relevant to investment companies that are aiming for an elevated level of ESG integration. These companies need to organize ESG scoring information, and they can have a motivation to generate their own ("shadow") ESG scores and to constantly monitor and audit ESG disclosure information—for example, to identify greenwashing and to close monitoring gaps. The scores are produced for ESG-integrating investment portfolios, such as by using cross-sectional systematic trading strategies to rank assets before portfolio construction. There is a constant tracking of trading portfolios, including AI-based detection of irregular transactions and behavior.

We are moving toward an evidence-based, data-driven, AI-powered ESG approach. If implemented correctly with the right data, data science approach, and IT/compute infrastructure, the improved analysis would be:

- neutral (in terms of no human bias), consistent, and less biased (in terms of unwanted bias);
- global, frequent, and available in a timely way;
- enabled in a bottom-up way; and
- comprehensive due to the coverage of undisclosed information.

Massive amounts of primary information such as text-based data (some of which are web-scraped) or alternative data from satellite imagery need to be collected, cleaned, streamed, and analyzed, activities that can be supported by machines very efficiently. On top of that are additional compute layers that help to trace back the processing steps and reasoning of the machines, which is important for human–machine interaction, such as providing validation, developing data narratives, and amplifying the subject matter experts. It would hardly be possible to reach those levels of reproducible and transparent information processing by manual analysis of those amounts of data.

Another example for innovative technologies that support the investment and risk management process is AI-fueled analytical capabilities, such as generating potential market scenarios that have never been observed before but can be viewed as realistic. In a second step, a machine learning (ML) program identifies links between properties of those market scenarios and the performance of a set of competing investment strategies. In the last step, a computationally intensive algorithm reveals the structures that the AI/ML system has learned in the training process. This step is particularly important because investment managers now have the option to understand the decision-making process of the AI/ML system using XAI to validate it and build a narrative for the choice of a certain investment strategy. This can be applied in many investment processes that seek to find the best way to diversify a given number of assets and aggregate them into a robust portfolio. In practice, this approach is applicable in a variety of investment companies, ranging from hedge funds to pension funds. The amount of computation involved is probably not feasible for a human being to execute manually in a lifetime.

Underrated Aspects of AI-Driven Investment Strategies

Successful investment firms of the future will be those that strategically plan to incorporate AI and big data techniques into their investment processes (Cao 2019). Besides the importance of advancements in mathematical modeling, several additional aspects of the AI strategy need to be addressed:

1. An appropriate infrastructure for AI training and inferencing, as well as for simulation for risk management, algorithmic trading, and back testing
2. A scalable enterprise-level workflow and process for developing and deploying AI models
3. Robust AI models that can be validated and explained, which removes barriers to AI adoption and builds confidence in these innovative technologies

All this is part of the skill set that investment companies can develop to make a difference and to use as a unique selling proposition.

A key technology to implement these three aspects is accelerated computing, which can be used for simulation, data manipulation, AI model building, and deployment/inferencing. Accelerated computing is widely available today, and it can increase developer team productivity, ROI (return on investment), model quality, and enterprise-level scalability while decreasing TOC (total cost of ownership), time to insight, energy consumption, and infrastructure complexity. It is all about building and using more effective models at higher speed and lower cost while keeping the models robust and explainable.

Data science and AI teams will be able to amplify their productivity using "AI and simulation factories." Using these is also a crucial factor in attracting talent that can do the "work of a lifetime" on such computing and accelerated data infrastructures.

The need for accelerated computing becomes noticeably clear when several data sources need to be acquired and combined. Several data sources are jointly aggregated, correlated, analyzed, and visualized in increasingly rapid ways. Technologies for accelerated data logistics and curation enable the gathering of datasets from dozens of sources—such as remote sensors, IoT (internet of things), social media, and satellite data. Iteration speed is increased by reducing analytical latency by seconds or even milliseconds. New forecasting techniques based on ML that use alternative, complex, unstructured datasets (such as satellite images) and new generative methods to create synthetic data are changing the way we produce and backtest strategies.

Investors around the world want to know more about the performance of individual companies before others do. They address leading indicators related to the impact of future financial results, sustainability, and environmental risk factors. Analysts need to effectively query and visually examine these vast datasets. Analytics software and AI/ML running on traditional CPU-based servers can be less capable of processing huge amounts of data required for advanced, multilayered, and multimodal analysis in a timely manner and might use more energy than graphics processing units (GPUs) do.

The Accelerated Computing Revolution and How It Enhances Investment Management

GPUs can be found in many compute clusters, ranging from supercomputing to public clouds and even enterprise data centers. They can accelerate the processing of huge amounts of structured and unstructured large datasets and execute large training and inferencing workloads. These ultra-fast processors are designed for massively accelerated training epochs, queries, complex image rendering, and interactive visualization. Combined with purpose-built analytics software, they deliver the kind of speed and zero-latency interactivity that professional investors need.

Investment companies of all shapes and sizes are taking advantage of this revolution in GPU analytics. GPU-accelerated platforms enable faster analytics to uncover more sophisticated investment and growth opportunities. GPU-powered models offer higher throughput with lower latency. As a result, more sophisticated models can be used for a given latency budget, leading to far more accurate results.

As the world generates increasingly new forms of data that provide meaningful signals for conclusions to be drawn about future business and market performance, accelerated computing platforms can become an increasingly valuable technology.

GPU-based acceleration technologies are universally applicable and help not only with data collection, preparation, and model building but also in the production phase, including model deployment, inferencing, and data visualization.

MLPerf[4] is a consortium of AI leaders from academia, research labs, and industry whose mission is to "build fair and useful benchmarks" that provide independent evaluations of training and inference performance for hardware, software, and services—all conducted under prescribed conditions. GPU-based systems shine in these benchmarks regularly, as can be seen in the MLPerf tables.

Equally important is the model validation step, where the model and data are made transparent and explainable to reconnect them to human intelligence, creativity, and

[4] For more information, go to the ML Commons webpage at https://mlcommons.org/en/#philosophy.

domain expertise—a step often forgotten or underrated. The explainability step can also be computationally demanding, which is the reason accelerated computing is a valuable resource, as it realizes significant speedups (Mitchell, Frank, and Holmes 2022).

To summarize, accelerated computing can build more model alternatives, with potentially higher accuracy and at the same time at lower cost and energy consumption and with greater flexibility. Such approaches as "fail fast, fail forward" can be implemented with accelerated computing, which can be viewed as a kind of "time machine" by speeding up the iterations required for model development.

Exhibit 1 shows some areas where GPU-accelerated computing supports investment (risk) managers, traders, and (AI/ML) quants.

Exhibit 1. Overview of GPU-Accelerated Domains plus Corresponding Tools and Software Packages for Implementation

Domains Enhanced by GPU-accelerated Computing	Descriptions of Useful Tools and Software Packages for Implementation
Loading and preprocessing enormous amounts of data in dataframe operations	• Open GPU Data Science with RAPIDS[a] suite of software libraries to execute end-to-end data science and analytics pipelines • Exposes GPU parallelism and high-bandwidth memory speed through Python interfaces • RAPIDS accelerator for Apache Spark supports the following steps: data acquisition, preprocessing, manipulation, data curation • RAPIDS library cuDF is related to Python library pandas
AI/ML/NLP and financial data science	• GPU-accelerated versions of PyTorch, TensorFlow, and other widely used deep learning frameworks • TAO: transfer learning based on optimized architectures and pretrained models • RAPIDS (as described before): end-to-end data science and analytics pipelines • RAPIDS cuML: GPU-accelerated versions of XGBoost and scikit-learn (like unsupervised learning, clustering decomposition, and dimensionality reduction); GPU-accelerated training for forest models, such as XGBoost, LightGBM, scikit-learn random forest • GPU-accelerated network analysis with RAPIDS cuGraph (linked to NetworkX) for community detection, link analysis/prediction, traversal, centrality, tree filtering, network visualization, Graph Neural Networks with Deep Graph Library (DGL), and PyTorch Geometric (PyG) • Dask[b] integrates with RAPIDS and can distribute data and computation over multiple GPUs, either in the same system or in a multinode cluster • MLOps (machine learning operations) tools, such as high-performance deep learning inference (such as a deep learning inference optimizer and runtime that deliver low latency and high throughput for inference applications) and inference server software (helps standardize model deployment and execution and delivers fast and scalable AI in production)[c] • Learning to rank (e.g., for constructing cross-sectional systematic strategies) • End-to-end frameworks for training and inferencing large language models (LLMs) with up to trillions of parameters[d] • Cloud-native suites of AI and data analytics software optimized for the development and deployment of AI, like an enterprise AI operating system for the accelerated AI platform[e]

(*continued*)

Exhibit 1. Overview of GPU-Accelerated Domains plus Corresponding Tools and Software Packages for Implementation (*continued*)

Domains Enhanced by GPU-accelerated Computing	Descriptions of Useful Tools and Software Packages for Implementation
Model risk management and human–machine interaction	• AI model validation and managing AI risk in the enterprise for AI models in production • Transparent, explainable, robust, auditable models for building trust techniques to test and improve robustness, explainability, fairness, and safety of the models, including large-scale visualization of data and model results with a usually considerable amount of compute • AI-generated synthetic data further enhance the model building process: Synthetic data improve model stability and explainability but can also generate stress-test scenarios that have never been observed but are realistic at the same time
High-performance computing (HPC) for derivative pricing, risk management, and portfolio allocation	• GPU-accelerated software development kits (SDKs) for Monte Carlo risk simulations for market risk applications (exotic derivative pricing, variable annuities, modeling underlying volatilities—e.g., Heston), counterparty risk (CVA, XVA, FVA, MVA valuation adjustments) and market generators/simulators • Relevant benchmark is STAC-A2[f] for risk models across throughput, performance, and scalability[g] • Advanced compilers, libraries, and software tools are available for real-world financial HPC applications, including American and exotic option pricing with multiple methods of writing GPU-accelerated algorithms, including the use of accelerated ISO (International Organization for Standardization) Standard C++ • Risk calculations with GPUs can be done 40 times faster, reducing costs by 80%. A Federal Reserve analysis showed speedups of over 250 times for GPUs vs. CPUs (central processing units) in running Monte Carlo simulations for European and American options pricing[h]
Algo trading, risk management, and backtesting	• Algorithm trading requires AI/ML workloads, filters, backtesting engines, bootstrapping scenarios, and optimization algorithms. Algorithm trading is a top use case with high growth rates and penetration. It makes use of HPC and GPU-accelerated deep learning for training and inference • GPU acceleration delivered more than 6,000 times speedup on the STAC-A3 benchmark algorithm for hedge funds[i]
Quantum computing with use cases in portfolio optimization, MC simulation, and ML	• GPU-powered speedup of quantum circuit simulations based on state vector and tensor network methods by orders of magnitude[j] • Hybrid quantum-classical computing involving QPU, GPU, CPU[k]

[a] www.nvidia.com/en-us/deep-learning-ai/software/rapids/.

[b] www.dask.org/.

[c] https://github.com/NVIDIA/TensorRT and https://github.com/triton-inference-server/server.

[d] https://github.com/NVIDIA/NeMo.

[e] www.nvidia.com/en-us/data-center/products/ai-enterprise/.

[f] www.stacresearch.com/a2.

[g] Some of the largest firms on Wall Street and the broader global financial industry rely on STAC-A2 as a key risk model benchmark to measure compute platform performance. "The STAC-A2 Benchmark suite is the industry standard for testing technology stacks used for compute-intensive analytic workloads involved in pricing and risk management" (www.stacresearch.com/a2). For example, it measures the time to complete the calculation of a set of Greek values for an option (which measure the sensitivity of the price of an option to changes, such as price of the underlying asset, volatility, or interest rates). Thus, Greeks—which should be recalculated as an option's price varies—provide a risk management tool for assessing market impacts on a portfolio of options.

[h] Scott (2018).

[i] Ashley (2019).

[j] https://github.com/NVIDIA/cuQuantum.

[k] https://developer.nvidia.com/cuda-quantum.

Further information can be found in Business Systems International's executive guide for the use of GPUs in data science and quantitative research in financial trading.[5]

Three Examples of Accelerated Computing in Real-World Investment Use Cases

This section demonstrates the GPU-accelerated tools and software package from the previous section in action. I describe three cases—two on ESG investing and one on diversified portfolio construction.

ESG and Risk Analysis Using NLP and Document Intelligence

NLP is one of the most prominent AI techniques that processes text information for many tasks. Among them are named entity recognition, topic modeling, intent identification, relation extraction, sentiment analysis, language translation, question answering, and text summarization.

Financial news and ESG disclosure reports exhibit complex formats along with images and tables, where document intelligence and AI are used to extract and digitize the relevant information (text, tables, pictures, etc.).

Sustainability reports need to comply with diverse taxonomies and reporting standards, which involves semantic understanding of various financial and nonfinancial disclosures. A central ESG and risk data repository can be built up to operationalize reporting and adapt to evolving regulatory requirements.

A real-time ESG analytics process drives ESG insights and scenario analysis with news and media screening. Adverse event monitoring further enables investors to stay in control of both reporting and assessment. Advanced NLP technologies monitor unstructured data in news and social media articles. The insights can be used in conjunction with the ESG scores from the rating agencies to provide a holistic view and thus improve decision making based on techniques.

Investors and fund managers use these assistive technologies to detect and mitigate ESG fraud, such as greenwashing. Fraud and anomaly detection platforms that monitor firm disclosures and are powered by a comprehensive greenwashing detection framework will drive investor confidence and flow of funds to truly green entities—one of the biggest challenges for ESG investors today.

NLP-based detection of fraudulent disclosures, news, and communications is the first line of defense against companies misrepresenting brown assets to solve their problem of stranded assets and loss of valuation. Quantitative and measurable data are needed to enable comparability. There can be real-time assessments of issuer operations.

In all these NLP-driven use cases, it becomes clear that large language models can be greatly beneficial from an engineering perspective because these models can be adapted to various tasks with a few shots with examples and prompt engineering (Gopani 2022).

Over the past few years, some leading software and solution companies have been established, making use of accelerated computing to be able to quickly process vast amounts of data. Some of them have developed no-code environments[6] and numerous AI-based services and data around NLP for ESG information.[7]

Earth Observation Data and Spatial Finance

Earth observation (EO) is the gathering of information about planet Earth's physical, chemical, and biological systems via remote sensing technologies, usually involving satellites carrying imaging devices, delivering reliable and repeat-coverage datasets.[8] EO is used to monitor and assess the status of and changes in the environment.

EO and remote sensing combined with ML have the potential to transform the availability of information in our financial system (Papenbrock, Ashley, and Schwendner 2021).

"Spatial finance" is the integration of geospatial data and analysis into financial services. It allows financial markets to better measure and manage climate-related and environmental risks, such as loss of biodiversity, threats to water quality, and deforestation. According to the EU Agency for the Space Programme (EUSPA 2022, p. 12), the insurance and finance segment will become the largest contributor to global EO revenues in 2031 (with EUR994 million and an 18.2% market share).

[5] https://media.bsi.uk.com/white-papers/Business_Systems_International_(BSI)_-_Nvidia_GPUs_in_Quantitative_Research_Executive_Guide.pdf.

[6] See, for example, https://accern.com/esg-investing and https://demo.softserveinc.com/esg-platform.

[7] See NVIDIA's on-demand webinar "How No-Code NLP Drives Fast and Accurate ESG Insights": https://info.nvidia.com/nlp-risk-management-esg-financial-services.html?ncid=so-link-135948-vt09&=&linkId=100000133473284&ondemandrgt=yes#cid=ix06_so-link_en-us.

[8] See the EU Science Hub's "Earth Observation" webpage: https://joint-research-centre.ec.europa.eu/scientific-activities-z/earth-observation_en#:~:text=Earth%20observation%20is%20the%20gathering,the%20natural%20and%20manmade%20environment.

EO and remote sensing data can be extremely complex to process. Processing massive amounts of EO data from multiple sources involves complex workflows and ETL (extract, transform, load) processes that first load and transform the data in several steps and then apply advanced AI models for computer vision and environmental monitoring. Typical steps are image formation, calibration and geospatial processing, data streaming, decompression, computer vision, model inference, scientific visualization, and (photorealistic, cinematic, 3D) rendering.

Also, the more granular and high-resolution the data, the more insight can be generated in finance and trading as individual assets can be observed and tracked. Sometimes the data need to be streamed or must be analyzed over time ("spatiotemporal"). The value creation for financial insights and trading sometimes also requires fusing multiple sensor sources to create a context. Imagery needs to be corrected, and features/patterns need to be extracted.

Accelerated computing can be a key technology in spatial finance to process EO and remote sensing data in a fast, reliable, cost-efficient, flexible, and energy-efficient way, and the models need to be of high quality and high resolution to build financial services, products, and signals for specific customers. Today it is possible to construct end-to-end (streaming) pipelines based on GPU acceleration, including processing of JPEG files, geolocation algorithms, spatial and trajectory computations, streaming, and computer vision, including clustering and graph analytics in RAPIDS.[9]

In the finance industry, the uptake of EO is becoming increasingly important in informing decision-making processes often before markets are affected. There are several concrete use cases in commodity trading where biodiversity and conservation across value chains can be analyzed. Trading of such products as metals, agricultural products, and energy (gas, oil, etc.) is being monitored and factored into value chains to make energy price predictions.

Investment managers are interested in identifying supply–demand balance changes. These are fed into trading models, such as in the oil market. Disruption of supply in hard and soft commodities by extreme weather, climate change, natural or other disasters/accidents, and pollution can be evaluated by processing EO data. Doing so usually includes an analysis of the history and assessment of the current physical situation. Regarding soft commodities (agricultural products), the users are usually interested in predictions on the yield rates of the next harvest as early as possible in the growing cycle. Knowing about crop shortfall in one region helps diversify and hedge early on.

Another use case is geospatial ESG.[10] It uses geospatial data to gain insight into a specific commercial asset, business, portfolio, or geographic area. It starts with pinpointing the location and defining the ownership of a commercial asset, such as a factory, mine, field, road, or retail property, known as asset data. Using various spatial approaches, it is then possible to assess the asset against observational data to gain an understanding of the initial and ongoing environmental impacts, as well as other relevant social and governance factors.

Geospatial ESG leverages geospatial AI using satellite imagery and sensor data to detect environmental or social parameter violations of companies and supply chains, such as biodiversity reduction, water source pollution, and greenhouse gas emissions. These detection and analysis models further enable investors to make more-informed decisions about a company separate from its voluntary disclosures or to monitor conformity with standards once established. There are enhanced risk assessments of ESG criteria on their investments, and access to real-time geospatial, environmental, and social data is a critical tool to enhance their evaluations.

Geospatial ESG helps validate issuer disclosures in cases where other data and information are unavailable, making ESG assessments more reliable. EO data are creating the foundation for predictability in ESG scoring/ESG performance. They help in an evolving regulatory landscape, and they are real-time, quantifiable, and measurable data for strengthening ESG analysis.

NLP and geospatial AI for sustainable finance are not only concerns of investment companies, banks, and insurance companies. Central banks around the world also are collaborating on greening the financial system and use accelerated data science, AI, and GeoAI (Papenbrock, Ashley, and Schwendner 2021) for sustainable finance.

Asset devaluation and long-term risks in ESG investing must be kept in mind. Climate risk is a critical ESG focus today. Potential infrastructure and property losses due to climate change affect organizations' long-term financial sustainability. Many investors examine a company's preparation assessment and capacity to forecast and respond to a variety of climate threats and environmental changes.

There are transition risks but also physical risks, such as extreme weather and record temperatures, that are now recognized as events that can be predicted and factored into financial planning. Droughts, floods, and wildfires are all examples of acute threats, but tropical illnesses due to rising temperatures and loss of biodiversity also fall in this

[9] Such as the cuSpatial library for GPU-accelerated spatial and trajectory data management and analytics: https://github.com/rapidsai/cuspatial.

[10] For more information, see https://wwf-sight.org/geospatial-esg; EUSPA (2023); European Space Agency (2022); Spatial Finance Initiative (2021); Patterson, Izquierdo, Tibaldeschi, Schmitt, Williams, Bessler, Wood, Spaeth, Fang, Shi, et al. (2022).

category. Additional risks can occur because of political instability and conflicts that climate change may generate, including supply chains and critical infrastructure that can be affected by extreme weather events.

Investors are exposed to such complex, evolving risks. Data, AI, simulation, and visualization will help investors better understand the risks and develop actions and mitigation strategies.

The adoption of such use cases will be further increased when the quality, availability, and granularity of observational climate and environmental data are enhanced. The models for predicting extreme weather events and climate change will further improve, especially becoming more granular and of a higher resolution. This outcome can be achieved by such technologies as physics-informed AI and accelerated supercomputing. An example is FourCastNet (Pathak, Subramanian, Harrington, Raja, Chattopadhyay, Mardani, Kurth, Hall, Li, Azizzadenesheli, et al. 2022). Frameworks exist for developing physics ML neural network models.[11] Visualization of simulated scenarios on a highly granular asset level is a technique that helps analysts evaluate the risks and impacts. There are platforms for building large-scale digital twin simulations of environmental processes and risks.[12]

An example workflow would involve the following components:

- (continuous, streaming) satellite imagery to model inference pipelines, such as capabilities for land/sea classification systems and GISs (geographic information systems);
- compute-powered, physics-informed neural networks and Fourier neural operators for simplification for fast predictive capability of climatic events; and
- 3D rendering and visualization to understand what is happening and see the what-if scenarios.

AI and Simulation for Diversified Portfolio Construction

There are several problems with the way modern portfolio theory (MPT) and portfolio diversification are often implemented in practice. One is backtest overfitting on the same historical data, and another is using quadratic optimization with noisy covariance estimates as inputs. In this section, I will describe approaches to mitigate these problems. The approaches are consistent with the Monte Carlo backtesting paradigm and could address the replication crisis to a certain extent. Synthetic datasets should be the preferred approach for developing tactical investment algorithms (López de Prado 2019). This approach helps investors deal with the unknown data-generating process. Using synthetic data is not new, and it is like generating realistic artificial landscapes and training industrial autonomous machines, such as robots and self-driving cars. Synthetic asset return data are a way to test AI-driven investment strategies.

Approaches to diversified portfolio construction do not always require quadratic optimization but can be solved heuristically to reduce realized risk and heavy portfolio turnover. AI can even help investors decide when to switch the approach in an interpretable way.

A lot of research has been conducted and much progress has been made on generating synthetic asset return data with typical stylized facts of asset returns using various AI-based techniques. Less focus has been placed on generating correlated returns with stylized empirical facts of asset return covariance matrices. This is important for pricing and risk management of correlation-dependent financial instruments and portfolios. Approaches using GANs (generative adversarial networks) generate realistic correlation scenarios (Marti 2020). An approach called "matrix evolutions" uses evolutionary multi-objective optimization, which can be implemented in parallel to benefit from accelerators (Papenbrock, Schwendner, Jaeger, and Krügel 2021). The idea is to address multiple stylized facts of empirical correlation matrices at the same time and join them in a multi-objective optimization problem. The user defines ranges to generate "populations" of correlation matrices that cover a variety and combination of stylized facts, on the one hand, and exhibit properties of correct correlation matrices, on the other hand. In this way, millions of realistic correlation matrices can be generated that have never been observed before. Approaches to robust investment portfolios can be tested against those populations of correlation matrix scenarios to test and analyze their behavior and robustness.

Based on these or other artificially generated asset return data, it is possible to formulate a program for ML that predicts the outperformance of a certain portfolio construction approach based on information about certain market regimes. In this way, an investment client could produce a desired portfolio of assets or asset classes, and the investment manager could respond with a recommendation for which portfolio construction method should be preferred for a certain market phase or regime to realize a very robust diversified portfolio.

The approach works as follows: Each generated market scenario is represented in terms of its stylized empirical

[11] For example, see the "NVIDIA Modulus" webpage: https://developer.nvidia.com/modulus.

[12] For example, see the "NVIDIA Omniverse" webpage: www.nvidia.com/en-us/omniverse/.

facts—across the portfolio assets and including their correlation properties. These are the input features to the ML program. The labels can be measured by the outperformance of a certain portfolio construction method in that specific market scenario. By generating millions of those scenarios, one gets millions of labeled learning datasets, which are the training material for the ML program. This can be done for several available portfolio construction methods, and in the end, the ML program can be used to recommend a more robust portfolio construction approach in a specific market phase. In addition, the machine can even produce explanations in terms such as reporting the driving properties of certain market phases leading to an outperformance of a certain portfolio construction method. This outcome is achieved with post hoc explainers such as SHAP (SHapley Additive exPlanations), which are computationally intensive and can be GPU-accelerated, and with the massively parallel exact calculation of SHAP scores for tree ensembles (Mitchell et al. 2022).[13]

The entire workflow is described in an article by Jaeger, Krügel, Marinelli, Papenbrock, and Schwendner (2021). The idea is to produce alternative portfolio construction mechanisms, such as hierarchical risk parity (HRP), that use graph theory and representation learning to construct the portfolio by cutting through the estimation noise. In many market environments and many investment universes, these approaches offer lower realized risk than can be achieved by explicit portfolio risk minimization algorithms. This sounds unreal but is achieved by the two-step approach of HRP: (1) matrix seriation and (2) recursive bisection. In this way, the natural taxonomy of assets is preserved, and the less robust matrix inversion step can be skipped.

As this procedure does not work in all market environments, an AI-driven program can identify in which constellations the new method can be used. It can also deliver some explanations using Shapley values, a method based on cooperative game theory. This explainability helps portfolio managers identify the correct approach to portfolio construction depending on investment universe and market state (Papenbrock and Schwendner 2015). Interactive dashboards allow the validation and audit of the AI models and the extraction of evidence-based, data-driven insights and narratives.

The implementation of such an approach requires several computing-intensive steps to generate market data, to train AI models, and to finally extract information about the inner decision making of the models to draw conclusions. A normal desktop PC would be able to produce reliable outcomes after several hours, whereas a small GPU cluster can produce results after a few minutes (Papenbrock 2021). The entire workflow can be extended by many steps—namely, generating synthetic multidimensional market data in a very flexible and convenient way of matrix evolutions or producing completely new ways of constructing portfolios, as in an article by Schwendner, Papenbrock, Jaeger, and Krügel (2021). The approach and workflow can even be used to tame crypto portfolios (Papenbrock, Schwendner, and Sandner 2021).

Conclusion and Next Steps

In this chapter, I discussed some of the latest investment technologies and provided useful tools to leverage recent developments in AI infrastructure and software development.

I demonstrated how these technologies give access to new and more stable sources of alpha, advanced risk management technologies, higher levels of customization, and better meet the requirements of investment clients.

Acceleration technologies support the entire workflow and data science process—from loading large files and curating dataframe operations to conducting model building and inferencing. They can even perform the model validation and explanation steps. This support enables a more interactive and data-centric approach to AI.

The use cases on ESG investing and robust portfolio construction have demonstrated how accelerated computing platforms help leverage both alternative, unstructured data (NLP, computer vision, document intelligence) and classical time-series data, such as asset returns.

Factories for complex data processing, AI, simulation, and visualization help the industry build resilient, sustainable, and profitable investment products/services in a customized, transparent, and explainable way.

References

Alarcon, N. 2020. "Accelerating Automated and Explainable Machine Learning with RAPIDS and NVIDIA GPUs." *Technical Blog*, NVIDIA DEVELOPER (17 November). https://developer.nvidia.com/blog/accelerating-automated-and-explainable-machine-learning-with-rapids/.

Ashley, J. 2019. "NVIDIA Delivers More than 6,000× Speedup on Key Algorithm for Hedge Funds." NVIDIA (13 May). https://blogs.nvidia.com/blog/2019/05/13/accelerated-backtesting-hedge-funds/?ncid=so-twi-nrcmflssx6-84884.

Cao, Larry. 2019. "AI Pioneers in Investment Management." CFA Institute. www.cfainstitute.org/-/media/documents/survey/AI-Pioneers-in-Investment-Management.pdf.

[13] For a related blog on explainable ML with acceleration, see Alarcon (2020).

European Space Agency. 2022. "Space for Green Finance: Use Cases and Commercial Opportunities" (November). https://commercialisation.esa.int/2023/01/market-trend-space-for-green-finance/.

EUSPA. 2022. "EO and GNSS Market Report." www.euspa.europa.eu/sites/default/files/uploads/euspa_market_report_2022.pdf.

EUSPA. 2023. "EU Space for Green Transformation" (25 January). www.euspa.europa.eu/newsroom/news/eu-space-helps-drive-green-transformation.

Gopani, A. 2022. "How NVIDIA Trains Large Language Models." AIM (23 March). https://analyticsindiamag.com/how-nvidia-trains-large-language-models/.

Jaeger, Markus, Stephan Krügel, Dimitri Marinelli, Jochen Papenbrock, and Peter Schwendner. 2021. "Interpretable Machine Learning for Diversified Portfolio Construction." *Journal of Financial Data Science* 3 (3): 31–51.

López de Prado, Marcos. 2019. "Tactical Investment Algorithms" (26 September). Available at https://ssrn.com/abstract=3459866.

Marti, G. 2020. "CORRGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks." *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: 8459–63.

Mitchell, R., E. Frank, and G. Holmes. 2022. "GPUTreeShap: Massively Parallel Exact Calculation of SHAP Scores for Tree Ensembles." *PeerJ Computer Science* 8 (5 April): e880.

Papenbrock, Jochen. 2021. "Accelerating Interpretable Machine Learning for Diversified Portfolio Construction." *Technical Blog*, NVIDIA DEVELOPER (29 September). https://developer.nvidia.com/blog/accelerating-interpretable-machine-learning-for-diversified-portfolio-construction/.

Papenbrock, Jochen, and Peter Schwendner. 2015. "Handling Risk-On/Risk-Off Dynamics with Correlation Regimes and Correlation Networks." *Financial Markets and Portfolio Management* 29: 125–47.

Papenbrock, Jochen, John Ashley, and Peter Schwendner. 2021. "Accelerated Data Science, AI and GeoAI for Sustainable Finance in Central Banking and Supervision." Paper presented at International Conference on Statistics for Sustainable Finance, Paris (September). www.bis.org/ifc/publ/ifcb56_23.pdf.

Papenbrock, Jochen, Peter Schwendner, Markus Jaeger, and Stephan Krügel. 2021. "Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios." *Journal of Financial Data Science* 3 (2): 51–69.

Papenbrock, Jochen, Peter Schwendner, and Philipp Sandner. 2021. "Can Adaptive Seriational Risk Parity Tame Crypto Portfolios?" Working paper (15 July).

Pathak, Jaideep, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. 2022. "FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators." Preprint, arXiv (24 February). https://arxiv.org/pdf/2202.11214v1.pdf.

Patterson, David, Pablo Izquierdo, Paolo Tibaldeschi, Susanne Schmitt, Alicia Williams, Janey Bessler, Steven Wood, Mike Spaeth, Fei Fang, Ryan Shi, et al. 2022. "The Biodiversity Data Puzzle: Exploring Geospatial Approaches to Gain Improved 'Biodiversity' Insight for Financial Sector Applications and the Pressing Need to Catalyze Efforts." WWF-UK (December). www.wwf.org.uk/sites/default/files/2022-12/The-Biodiversity-Data-Puzzle.pdf.

Schwendner, Peter, Jochen Papenbrock, Markus Jaeger, and Stephan Krügel. 2021. "Adaptive Seriational Risk Parity and Other Extensions for Heuristic Portfolio Construction Using Machine Learning and Graph Theory." *Journal of Financial Data Science* 3 (4): 65–83.

Scott, L. 2018. "Finance—Parallel Processing for Derivative Pricing" (March). www.nvidia.com/en-us/on-demand/session/gtcsiliconvalley2018-s8123/.

Spatial Finance Initiative. 2021. "State and Trends of Spatial Finance 2021: Next Generation Climate and Environmental Analytics for Resilient Finance." www.cgfi.ac.uk/wp-content/uploads/2021/07/SpatialFinance_Report.pdf.

11. SYMBOLIC AI: A CASE STUDY

Huib Vaessen
Head of Research and Analytics Real Assets, APG Asset Management

The degree to which machines can help in investment decision making varies widely, depending on the investment strategy. The solution discussed in this case study is based on symbolic artificial intelligence (AI) and serves investment teams using fundamental investment strategies.

Samuel: An Automated Real Estate Portfolio Management Solution Based on Symbolic AI

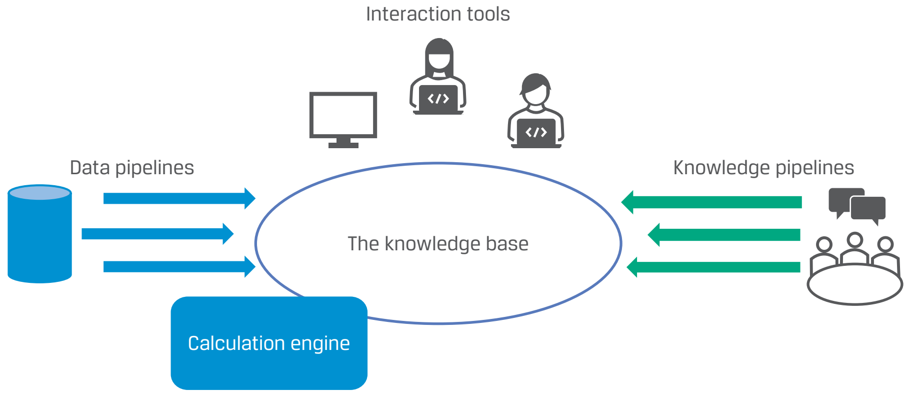
Samuel is a composite AI[14] system that collaborates with humans. It acts as a digital colleague that guides the human investment team with transparent, systematized, and well-substantiated advice. Its decisions are transparently substantiated and can be tracked down to each datapoint used, allowing for efficient reconciliation with the thought process of the human team.

A composite AI system uses the AI techniques most suitable for the result. Although nonsymbolic AI, such as machine learning (ML) and its subclass neural networks, has gained much attention, many real life use cases warrant the use of symbolic AI techniques. Symbolic AI, or classical AI, is a collection of techniques that are based on human-readable and high-level representations of the problem. Examples are rule-based engines and decision trees. It was the dominant approach to AI research since its early days through the mid-1990s (Russell and Norvig 2020).

First, the use case of an automated portfolio management solution for a fundamental investor warrants typical symbolic AI techniques because the amount of available data on investments and markets has increased but is still not sufficient to allow for ML techniques in many cases. Second, given that the interaction between human and machine is crucial—because humans have to be able to understand the reasoning for high-stakes investment decisions—the transparency and interpretability of the technique is important. Third, typically there is already important human knowledge available in a fundamental investment team that one might want to leverage. Symbolic techniques allow for codifying this human knowledge as one can imagine with rule-based systems. Many systems built with symbolic AI techniques are called "expert systems." In this chapter, mainly systems built with symbolic methods are described.

Technically, the automated solution can have many forms, but in general, it constitutes the following parts as depicted in **Exhibit 1**: a knowledge base, data pipelines, knowledge pipelines, the calculation engine, and tools such as dashboards that interact with the human portfolio manager.

Exhibit 1. Example: APG Asset Management Real Estate (APG RE)



Source: APG Asset Management.

[14] As defined in Gartner's "Hype Cycle for Emerging Technologies, 2020," composite AI refers to a "combination of different AI techniques to achieve the best result."

Next, I will discuss each component of Samuel in more detail.

The *knowledge base* is a database that contains all data related to the decision-making process and includes all principles that need to be applied to those data in order to get the outcomes. The knowledge base is fed by data pipelines for the data and knowledge pipelines for the principles. Humans are needed to provide oversight. The knowledge base can be stored in all kinds of relational and graph databases.

The *calculation engine* applies the rules to the data and stores the outputs in the knowledge base. The calculation engine requires human oversight and needs to be configured by humans. It contains a "worker" with CPU (central processing unit) power that does the calculations and a trigger mechanism that triggers the calculations when needed.

Interaction tools enable the interaction between Samuel and humans. The output of Samuel needs to be interpreted by humans, and the input needs to be given by humans. Interaction tools can have many different forms, ranging from dashboards accessible via the web browser to search bars and forms for input. If the interaction between the humans and the digital colleague is not well organized, the hurdles to provide input will be too high and the output less valuable. Similarly, if the output cannot be found by the right people at the right time, the impact of Samuel is lost. To ensure appropriate interaction tools, constant adoption is needed to embed Samuel in the other business processes to minimize the hurdles to use the output and to provide input by humans.

The *data pipelines* are ETL (extract, transform, load)[15] flows that ingest data from various sources outside the team. Here the IDs are matched, and data are prepared for further usage. The formatting of the sources changes over time, and data pipelines might have to be adjusted for that. Additionally, new data pipelines need to be added over time and integrated with the other data. Humans are needed to build the pipelines and provide oversight. Data pipelines can be written in Python, C++, or any other coding language and triggered by an orchestration tool that decides when to run the pipeline.

The *knowledge pipelines* are an active collaboration between humans and Samuel. Knowledge is constantly evolving, and as such, the principles need to be updated regularly. Thus, it is important that whenever new knowledge is created after a group discussion, this knowledge is made explicit and codified in the form of principles. Making knowledge explicit is a time-consuming process because often implicit knowledge is incomplete or inconsistent over persons or situations. Before a team has a set of principles that adequately describes the knowledge that is consistent among persons and situations, many discussions must have been held.

Samuel: APG Real Estate Investment Team's Digital Colleague

The APG real estate investment decision-making process is top down—first an allocation across regions, then an allocation across peer groups (sectors) within regions, and lastly the selection of the opportunities within each peer group.

The real estate team introduced Samuel in the winter of 2021.

In the initial phase, a team has to map its decision-making process, state which arguments are used, and explain how these arguments are substantiated. This part of the process is a project that involves many discussions between the members of the investment team to agree on the relevance of each argument and how to substantiate it. This project takes time but is a one-off. After that, refinements of the decision-making process are an ongoing add-on so are less intense.

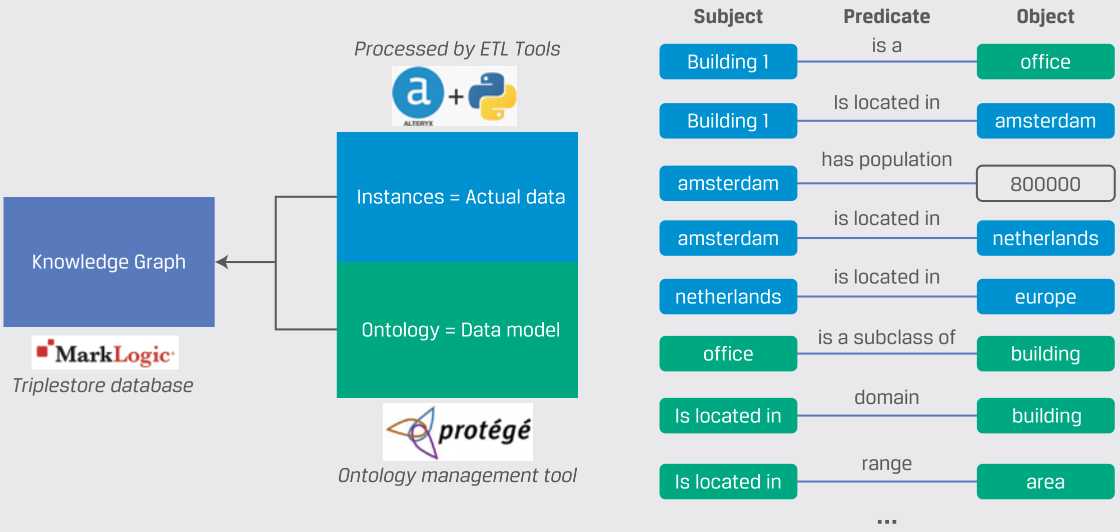
[15] *ETL* is a typical term used to indicate a process that extracts data from one source, transforms it, and loads it into another source.

Samuel collects all relevant input and updates the model when there are new data available. All this input is available for the portfolio managers' reference and forms the baseline for portfolio decisions. Whenever a human proposal is made, Samuel can also challenge that decision.

For example, Samuel updates the rental growth variable in the discounted cash flow model based on preset rules. Office rental growth is based on the quality of the building, which, in turn, is determined by the employment growth of the city, sustainability index of the building, and supply constraint in the market.

At APG Asset Management, the data that are stored in Samuel's brain are stored in a graph database as so-called triples, meaning that each fact is stored as a relation in the form of a subject–predicate–object triple, where the predicate defines the relation between the subject and the object. **Exhibit 2** shows an example of triples and the technology stack.

Exhibit 2. Examples of Triples and the Technology Stack



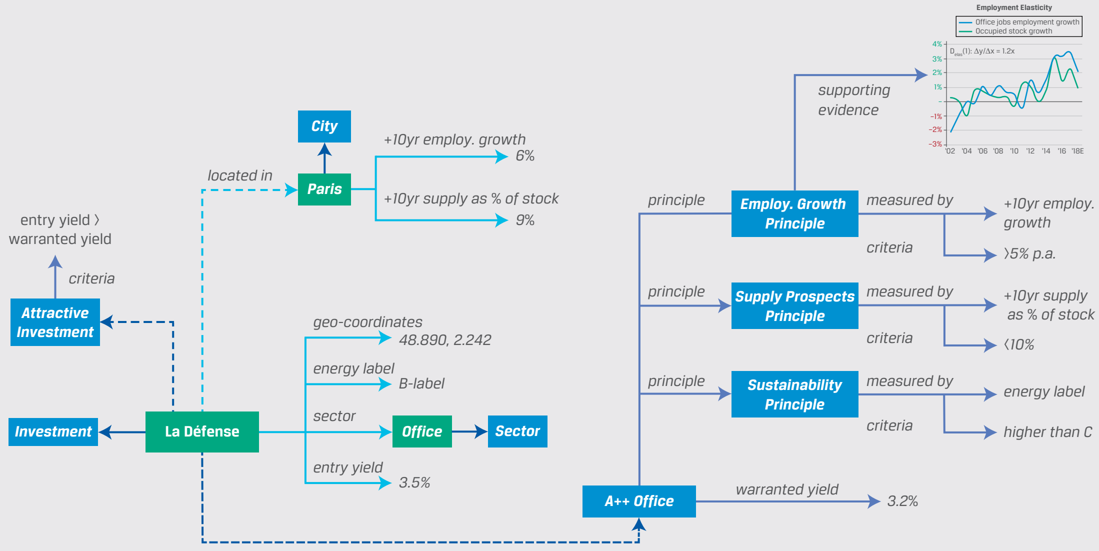
Subject	Predicate	Object
Building 1	is a	office
Building 1	Is located in	amsterdam
amsterdam	has population	800000
amsterdam	is located in	netherlands
netherlands	is located in	europe
office	is a subclass of	building
Is located in	domain	building
Is located in	range	area
	...	

Source: APG Asset Management.

Based on the relations shown in Exhibit 2, Samuel can infer by itself additional triples, such as the number of buildings owned by a certain company in Amsterdam, the Netherlands, or Europe. **Exhibit 3** shows a more complex example.

Storing information as triples with their relations allows for advanced questions where part of the answer is reasoned by a reasoner applied on top of the graph database. The following are examples of advanced questions: What are the principles that determine the quality of an office? What is the supporting evidence that employment growth is important for the quality of an office asset? What are the supporting arguments for the attractiveness of the office building "La Défense"? The triples are stored according to W3C standards, which makes retrieving and inference possible because you can use generic tools that work on these standards.

Exhibit 3. More Complex Example of Inferred Triples



Source: APG Asset Management.

The advantages of using a graph database to store the information are that it is easy to add new types of data, one can derive new knowledge based on the existing triples, and all data and metadata are stored in the same repository (which makes data more searchable while also allowing for a rich description of the data for future interpretation) and a graph database (a database that does not consist of multiple tables that need to be joined). Both structured and unstructured data can be joined in the same database, as can be seen in Exhibit 3, where the supporting document for the employment growth principle is also stored in the same graph database.

In the example in Exhibit 3, the calculation engine is Python scripts that are run on a server. A scheduled job imports, for example, the employment growth per city via an application programming interface (API) from an external provider. The graph database identifies this update as a trigger to start inferencing whether "La Défense" still is an A++ quality building given that the employment growth outlook for Paris has changed.

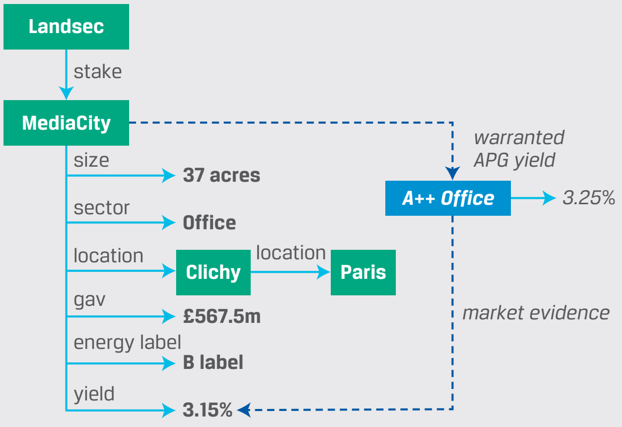
In the example, the main interaction tool is the cloud-based discounted cash flow (DCF) model. This model is deeply embedded in the day-to-day workflow of the human portfolio manager. In this cloud-based DCF model, Samuel can provide its suggestions for the rental growth as a field right next to the rental growth input field. As such, whenever the portfolio manager (PM) needs to enter a growth rate assumption, Samuel's suggested value is always at hand.

Exhibit 4 provides an example of how a news email about a real estate transaction can be read with natural language processing techniques and translated into triples for the graph database. The news item reads that the company Landsec acquired a stake in a building called MediaCity. The details of the deal can be recorded as triples and serve as input—for example, the quality assessment of the portfolio of buildings of the company Landsec.

Exhibit 4. Example News Email

Kempen Daily

Landsec today announces that it has acquired a majority stake in MediaCity, a 37 acre media, digital and tech office hub in Clichy, previously owned by a 50:50 JV between Legal & General and Peel L&P. Landsec will partner with Peel L&P who will retain a 25% stake and continue to serve as asset and development manager. MediaCity has a gross asset value of £567.5m. The location hosts tenants such as BBC North and ITV to Ericsson, The Hut Group, Kellogg's and over 250 creative and tech businesses as well as schools and universities. It is home to 8,000 residents. Landsec said that the scheme was renovated 10 years ago to improve its energy efficiency label to B. It is 96% let with a WAULT of just under 10 years. It generates £31.1m of net operating income per annum (100%), reflecting a 3.15% net initial yield.



Source: APG Asset Management.

How Do PMs Interact with Samuel?

Interaction with Samuel during the investment process can be grouped into four types: evaluating the buy/hold/sell decision, preparing the proposal, portfolio monitoring, and letting Samuel learn.

The evaluation of the final decision is the moment where the human-proposed action is ranked by Samuel versus all other potential actions. If other actions are more favorable, the portfolio managers that do the proposal will have to explain why they deviate from Samuel.

In preparation of the investment proposal, the human portfolio manager can be helped by Samuel with contextualized information. For example, while making the DCF model, Samuel can provide default values. The human PM can decide whether to overwrite the assumption or stick with Samuel's assumption.

Compared to the human team, Samuel excels in portfolio monitoring. Samuel collects the characteristics to monitor from each individual investment and contains the portfolio construction principles. As an example, each investment is mapped to countries and sectors, all stored in the knowledge base, complemented by the maximum weights that are desired in a country or sector. Samuel can check every day whether the aggregated investment exposure to a country or sector exceeds the limits. This repetitive work fits Samuel well because it comes at limited cost to schedule a daily execution of the check. Variables to monitor, such as the valuations of investments, leverage, and the beta of the portfolio, can be added at a low cost. The more variables of the investment portfolio are monitored, the bigger the difference between human and Samuel's execution in terms of consistency and effort.

The last interaction discussed here is a flow of information from the human portfolio manager to Samuel. Whenever after reconciliation of the proposed action with Samuel a consistent omission in Samuel's reasoning is found, its principles need to be adjusted. Doing so requires that the omission be made explicit, and it has to be made clear what new principles need to be added. This adjustment requires a well-facilitated discussion and the embedding of the new principles and potential new data sources in Samuel. This process needs to be guided. The continuous learning aspect is critical. Samuel learns by adding new rules and new data pipelines that provide the data on which the rules are applied. In this way, the rankings become more and more sophisticated and the default values harder to ignore.

Use Case APG Real Estate (continued)

Within the real estate team at APG Asset Management, Samuel is embedded in the investment process in several ways. The global team covers an investment universe of 300 listed real estate companies and sources hundreds of private investment opportunities annually. As such, there is a high volume of investment proposals that need to be evaluated.

Each investment proposal is discussed by the team in an initial phase. A subteam of the global real estate team is tasked with interpreting the outcomes of Samuel and challenging the proposal on the basis of Samuel's input during the discussion. After this initial screening, all attractive propositions are worked out in more detail, and then they are presented to the Real Estate Investment Committee. The output of the digital colleague also is presented to the committee.

While building the proposal, the portfolio manager is helped by contextualized suggestions embedded in the workflow. The cash flow model is cloud based and transparent for all members of the investment team globally. Samuel interacts with the human portfolio manager by providing contextualized information within the model in the form of references, such as growth expectations from external research providers, or default values, such as data-generated rental growth forecasts and discount rates. In our example of rental growth, the PM sees a proposed rental growth generated by Samuel. The PM can click through via the web browser to the components. Then, each component, such as the quality rating, is further substantiated.

Another argument, besides the output of the DCF, is, for example, portfolio fit: How does this fit in the top-down vision on sector and countries, and how does this contribute to environmental, social, and governance (ESG) goals? Samuel provides an automatic assessment of these factors to be used in the proposal. Whenever there are reasons to deviate, these can be addressed in the proposal.

How Does Samuel Improve Results?

Building and maintaining an automated system such as Samuel requires attention to the investment process on a higher abstraction level than a typical portfolio manager is used to. It demands attention for when and which information is needed for which decision. It also involves skills that are not traditionally available in a fundamental investment team, such as ontology building, data engineering, and software engineering.

These skills can be embedded in the organization in various ways, ranging from being centrally located to being located within the investment team (T-shaped teams) or within each investment professional (T-shaped skills). T-shaped teams are crucial for the development and maintenance of a recommender system such as Samuel. A T-shaped team is a team that contains individuals with the traditional investment skill sets, those with the data- and technology-related skill sets, and those with a mix of these skill sets (CFA Institute 2021).

The advantages of collaborating with a digital colleague are threefold. First, it improves decision making by countering human cognitive biases. A simple rule-based model tends to outperform human judgment (Kahneman, Sibony, and Sunstein 2021) because cognitive biases cause bias in decision making. By codifying best practices and applying them consistently, Samuel improves decision quality. There is usually unwanted variation in human decision making that is not warranted by the facts. This noise can be across persons or from the same person at different times. Experts typically believe they are themselves consistent and believe that other experts would decide in a similar fashion as themselves. However, research by Kahneman et al. (2021) shows that, for example, underwriting experts who are presented the same case have material differences in their judgments. Samuel makes judgments based on agreed-on principles and guidelines and, as such, provides a consistent benchmark.

The second advantage is that it is more transparent. Transparency in decision making allows the investment team to explain clearly to clients and stakeholders how their decision was formed. For example, you can track how responsible investment considerations affect your decision making.

The third advantage is improved efficiency of the decision-making process and results. The final decision is supported by arguments, and these arguments are the result of many decisions on a lower level that are typically less complex and made more frequently, some of which can be externally sourced and provided via Samuel. Another example of a lower-level decision that can be automated to a high degree would be the building blocks for the discount rate that is used. Many of these can be automatically updated by Samuel.

Although Samuel is a valuable addition to any investment team, it is also worthwhile to discuss the edge of the human portfolio manager over Samuel and show why a collaboration is so important.

First, humans can form a vision of the future and adjust their assumptions toward that future. Samuel relies on principles that were formed in the past (by humans) and might not yet have been adjusted to a new reality, such as in March 2020 when COVID-19 was affecting the world. A human portfolio manager knows when old principles are not applicable anymore given that the circumstances have changed.

Second, Samuel likely cannot operate on its own. Although parts of arguments can be input automatically with data from external sources combined with predefined principles, there are also many assumptions for which you need a human eye. Samuel uses the human assumptions where needed and takes its own assumptions if data and principles are available. In that sense, the output of Samuel is also based on human input.

Lastly, there are likely edge cases in which the systematic way of thinking does not translate well. These cases must be recognized by humans. In the end, Samuel bases its decision on a simplified model of reality.

Conclusion

The role of machines in the decision-making process will grow. Having a digital portfolio manager on the team can improve the quality of decision making, transparency, and efficiency. To build and maintain an automated digital portfolio management solution, it is necessary to embed nontraditional investment skill sets, such as ontology building, data engineering, and software engineering, within the investment department.

References

CFA Institute. 2021. "T-Shaped Teams: Organizing to Adopt AI and Big Data at Investment Firms." www.cfainstitute.org/-/media/documents/article/industry-research/t-shaped-teams.pdf.

Kahneman, Daniel, Olivier Sibony, and Cass Sunstein. 2021. *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark.

Russell, Stuart, and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach*, 4th US ed. Hoboken, NJ: Prentice Hall.